This model solution is provided so that candidates may better prepare for future sittings of Exam PA. It includes both a sample solution, in plain text, and commentary from those grading the exam, in italics. In many cases there is a range of fully satisfactory approaches. This solution presents one such approach, with commentary on some alternatives, but there are valid alternatives not discussed here.

Exam PA December 8, 2020 Project Report Template

Instructions to Candidates: Please remember to avoid using your own name within this document or when naming your file. There is no limit on page count.

Also be sure all the documents you are working on have December 8 attached.

As indicated in the instructions, work on each task should be presented in the designated section for that task.

Task 1 – Explore the variables (7 points)

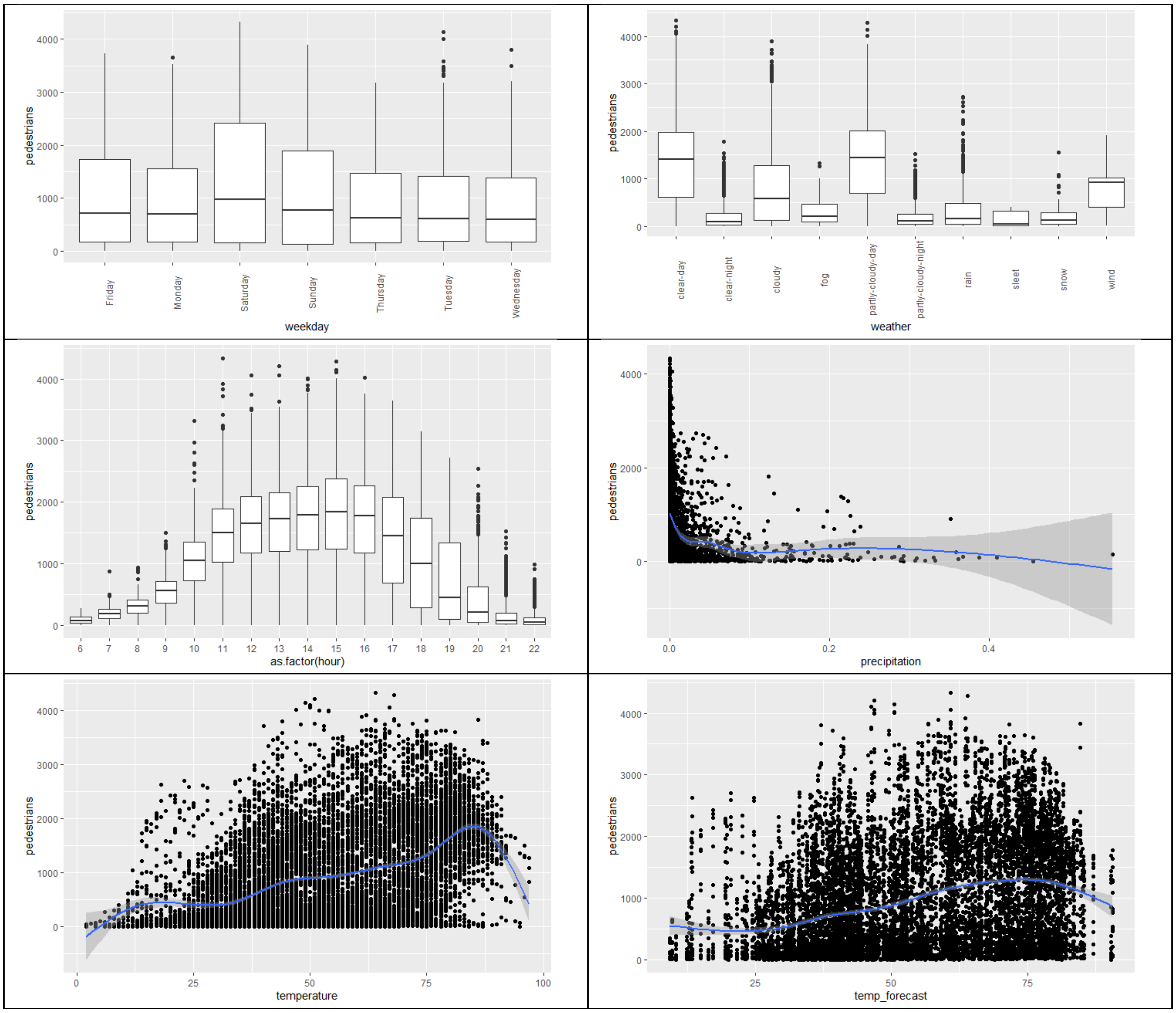
Some candidates showed plots and described relationships for all variables, while others only did that for the most and the least predictive variables. Either approach could earn full points as long as it was coupled with adequate justification. Many candidates failed to clearly identify a single variable they expected to be the most predictive and a single variable they expected to be the least predictive. Some candidates limited their discussion to analysis of graphs and did not identify important issues with the variables. For example, analysis of the weather variable often discussed the different levels of pedestrian traffic associated with each category, but it rarely discussed the poorly defined categories themselves.

The use of bold for identifying variable names when they are common words is not required but can help clarify the writing.

The following plots illustrate the relationship of each variable to the target variable (**pedestrians**). Each variable on its own appears to have a useful relationship with the number of pedestrians. Ultimately, a model will be needed to untangle the many variable interactions and determine the most and least useful features, but we can hypothesize based on visually inspecting the plots and looking at summary statistics. In order for the variables to be useful, the variable values will need to be able to separate observations that have high pedestrian traffic from those with low pedestrian traffic. I'll compare the medians from the boxplots and the ranges of the loess curves to assess that capability.

Comparing the medians of the boxplots: **hour** has the largest difference between median pedestrian values among its boxplots (1842 pedestrians at **hour** = 15 compared to 45 pedestrians at **hour** 22 = a difference of 1797), the largest observed difference for **weather** was 1387, and the largest observed difference for **weekday** was 380.

Comparing the maximum and minimum values from the loess curves: the difference between the maximum loess curve value and minimum loess curve value for **precipitation** is under 1000, the difference for **temperature** appears to be approximately 1800, and the difference for **temp_forecast** appears to be approximately 900.



Temperature

The **temperature** variable is least likely to have a significant contribution to the model. There are two variables that yield very similar information, **temperature** and **temp_forecast**. They have a very strong positive correlation (approximately 0.97), and we should not keep both variables in our final model, so one of these is the least likely to contribute since it won't be used. **Temperature** appears to have a stronger relationship to the target variable, but temperature typically increases throughout the day as the sun comes up and declines as the sun sets, and that will make it difficult for our models to separate the effects of **hour** and **temperature**. In fact, it is possible that the strong relationship seen for **temperature** above is driven more by the hour than the temperature. For this reason, I think it is more likely that I will remove that variable, so it will not contribute to the model at all.

Hour

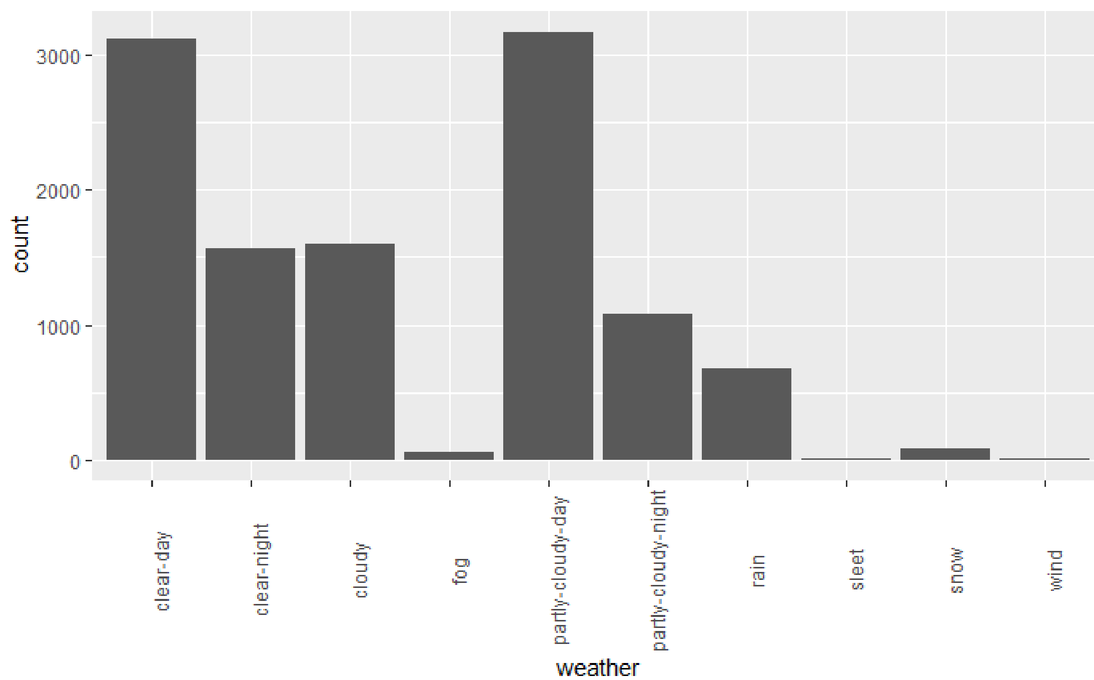
The **hour** variable is most likely to have a significant contribution to the model. Every value of **hour** is well represented in the dataset. The side-by-side boxplots above for **hour** illustrate that the number of pedestrians increases at the beginning of the day, peaks mid-day (the highest median pedestrians was at hour 15), and then declines in the later portion of the day. This is unsurprising since many people are asleep during the earliest hours and walking after dark may not appeal to many people.

Task 2 – Reduce factor levels (7 points)

Many candidates successfully identified levels that should be combined based on having a small number of records. Stronger candidates also identified other reasons to combine levels (e.g. how the levels of the variable related to the business problem, target variable, or other variables in the dataset).

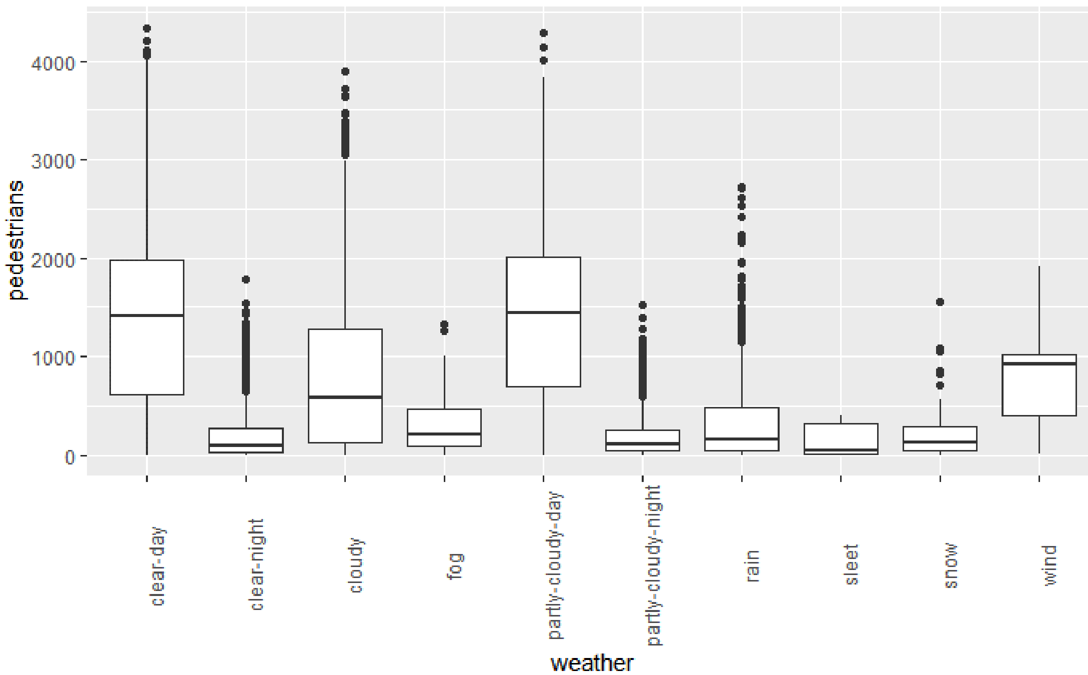
There are a number of issues with the **weather** variable:

- Some levels (fog, sleet, snow, and wind) contain very small numbers of records.



- Some categories are seemingly equivalent as far as the decision to walk or not is concerned (e.g., if it is nighttime, then does it matter if it is clear or partly cloudy?).
- The **hour** variable won't capture all of the daylight information because the sunrise and sunset times will vary throughout the year, but in this dataset **hour** appears to capture most of the daylight information (there is only a small overlap of day and night at some hours in the morning and evening), so having time information (day vs night) occasionally mixed in with the weather information is not very valuable.

The levels can be combined based on intuition about how the weather condition might affect the decision to walk, the effect of the hour variable explaining day vs night information, and whether the boxplots show a similar relationship to the target variable.



I will use the mapping below, removing the time component from **weather** and grouping mild and more inclement weather. Even though fog and rain have similar median pedestrian counts above, other factors may be interfering (e.g. fog typically only occurs at certain times of day) and it is better to rely on an intuitive sense of what weather is more or less pleasant to walk in.

Old levels	**New level**
“clear-day”, “clear-night”, “partly-cloudy-day”, “partly-cloudy-night”	“nice”
“wind”, “cloudy”, “fog”	“mild”
“rain”, “sleet”, “snow”	“inclement”

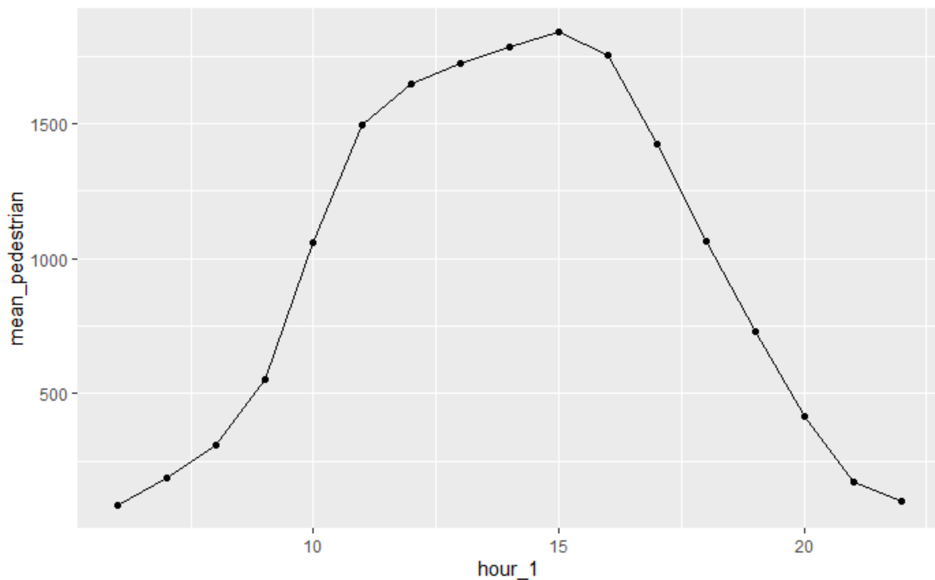
Task 3 – Modify the hour and temperature variables (11 points)

Most candidates performed well on this question. Other choices for the hour variable could have earned full credit as long as they were justified. For example, a justification for using the hour_1 version for the decision tree could have pointed out that the relationship for the hour variable is not perfectly symmetric, so using hour_3 will result in some information loss.

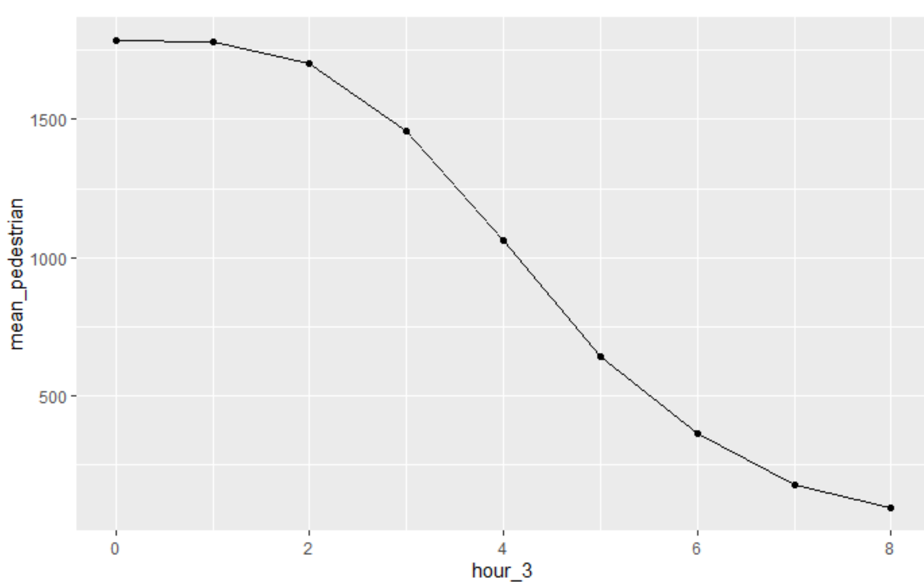
Lower performing candidates relied on justifications that demonstrated only a surface level understanding of the effects the choices would have on their models, such as stating that a decision tree performs well with factor variables as their sole rationale for their selection. Some candidates incorrectly stated that the hour_3 transformation created a linear relationship, or struggled to articulate the considerations around the two temperature variables (e.g., the *correlation of temp_1 with the hour variable). Stronger candidates included discussion relating the variables to the business problem in addition to their discussion of model mechanics.*

Hour

The original **hour** variable is the same as **hour_1**. The relationship to the number of pedestrians based on means is shown below.



The relationship to the number of pedestrians resembles a downward opening parabola. A similar relationship will exist for the **hour_2** variable with the difference that the models will be dealing with a factor variable instead of a numeric variable. The **hour_3** variable alters this relationship so that the mean pedestrians is monotonically decreasing as seen below.



Considerations for the Decision tree choice of hour variable

1. A numeric variable will result in easier to interpret splits, which is important for this business problem.
2. A factor variable could lead to a better fit to the signal in the data, but it is also more likely to overfit the data due to the specificity of the hour levels.
3. The **hour_3** version creates a monotonically decreasing relationship to the target variable. With binary splits in the decision tree, fewer splits may be needed vs **hour_1**, resulting in a simpler tree.

Based on the above considerations, I will use **hour_3** for the decision tree because it provides an opportunity for a good fit while maintaining interpretability and limiting the likelihood of overfitting.

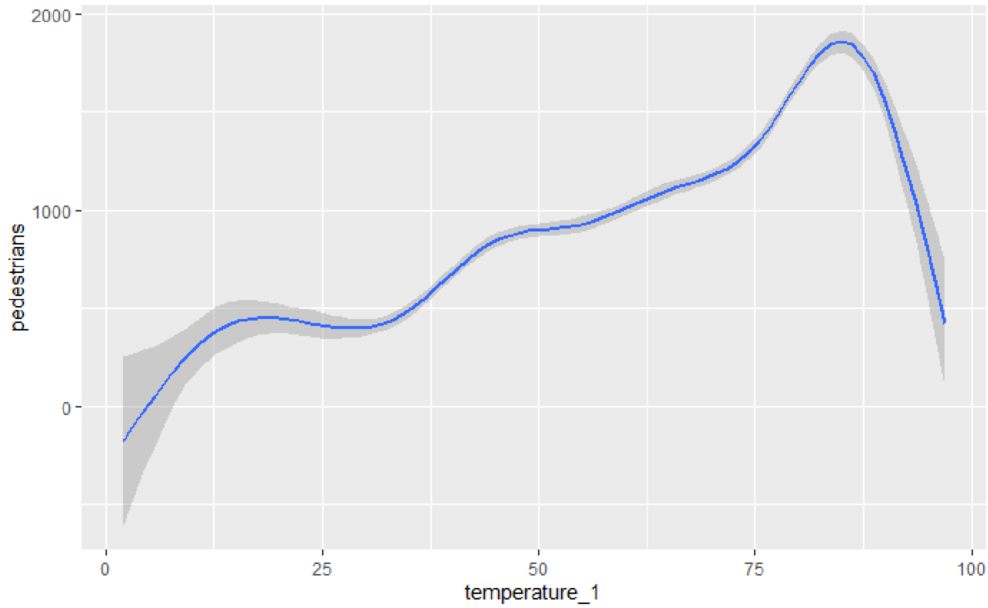
Considerations for the GLM choice of hour variable

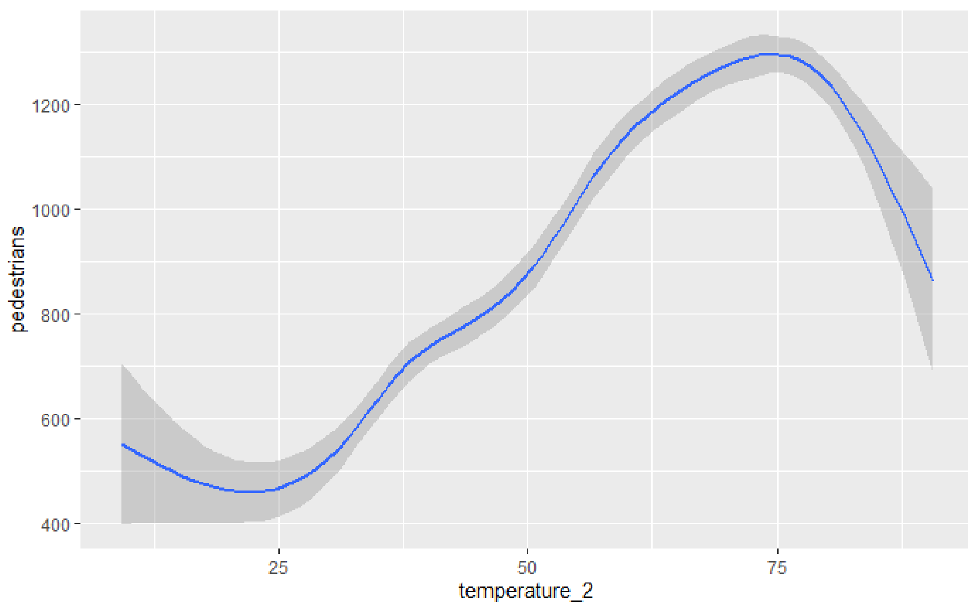
1. A numeric variable leads to a smaller (simpler) GLM that will be easier to interpret, which is a requirement for this business problem.
2. A factor variable could lead to a better fit to the signal in the data, but it is also more likely to overfit the data due to the specificity of the hour levels.
3. The **hour_3** version creates a monotonically decreasing relationship to the target variable and should produce a better fit without transformation than **hour_1**. This is because the downward parabolic shaped relationship to the target variable in hour_1 will require additional transformations (such as polynomial transformations) to fit well in a GLM. While that is an appropriate technique, it will result in a model that is difficult to interpret compared to a model built with **hour_3**.

Based on the above considerations, I will use **hour_3** for the GLM because it will provide a simpler GLM compared to **hour_2** and a better opportunity to fit without hard to interpret transformations than **hour_1**.

<u>**Temperature**</u>

The difference between the two versions of temperature is that **temperature_1** reflects the temperature changes throughout the day, while **temperature_2** summarizes each day's temperature with a single number. The loess plots below show that both higher values of **temperature_1** and **temperature_2** are generally associated with higher numbers of pedestrians. Both variables have similarly shaped relationships to the target variable (see next page).





Considerations for which temperature variable to use for both models

- It makes sense that people would not want to walk around outside when the temperature is too cold or too hot. The **temperature_1** variable does contain more information about the temperature, so as a stand-alone variable, it should be better at predicting the number of pedestrians.
- Even though **temperature_1** contains more information about the temperature, it is doubtful that people are changing their decision about whether or not they want to walk around throughout the day as temperature changes occur. It may be more common to plan whether or not they want to walk or not based on how warm it is supposed to be that day (summarized similarly to **temperature_2**). For the retail firm, it is even more doubtful that adjustments to operations will be made throughout the day based on the hour-by-hour temperature.

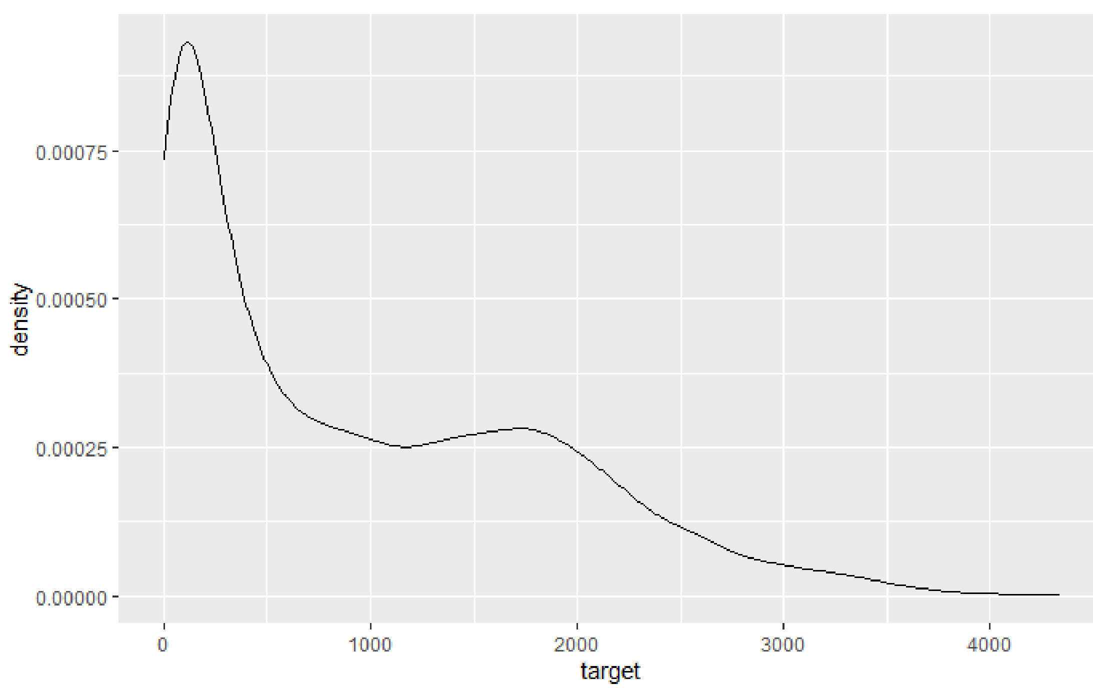
- There is a modeling challenge with **temperature_1**. It is correlated with hour because temperatures typically increase until the middle of the day before declining in the evening. It will be difficult for the models to separate these effects. Including both **temperature_1** and **hour** in a GLM will make the coefficients individually less interpretable because part of the impact on pedestrian traffic is explained by the other variable. Likewise, for a decision tree, any splits for **hour** or **temperature_1** must be interpreted by considering preceding splits of the other variable making any individual split less interpretable. The **temperature_2** variable does not suffer from this same problem, but it still yields useful information about the temperature.

Both versions of the temperature variable are related to the target variable and have an intuitive relationship to the number of pedestrians, but **temperature_2** is the better choice for our models because it avoids a complex relationship with **hour**.

Task 4 – Consider transformations of the target variable (9 points)

Many candidates incorrectly stated that non-linear transformations on the target variable do not have an impact on trees because trees are able to model non-linear relationships. Although it is demonstrated below, candidates were not expected to create multiple trees or perform any calculations to illustrate the impact of different transformations on the decision tree. Rather, a clear explanation of how transforming the skewed variable affects the tree was sufficient.

The original target variable is right skewed as seen in the density plot below.



Transformation impacts to the decision tree

With a right-skewed distribution, higher values will have greater influence over our decision tree model because splits are chosen that minimize the sum of squared errors. For each group created by a split, the sum of squared errors adds up the squared differences between the observed pedestrians and the mean pedestrians for the group. In a right-skewed distribution, the differences tend to be larger in magnitude for higher values than lower values, so they contribute more to the sum. Transforming the target variable impacts the sum of squared errors calculation, so the following items that rely on that calculation are also affected:

1. The location of splits
2. The number of observations in each leaf
3. The predicted values for each leaf – with a right-skewed distribution the predictions tend to be higher than they would be with a less-skewed distribution

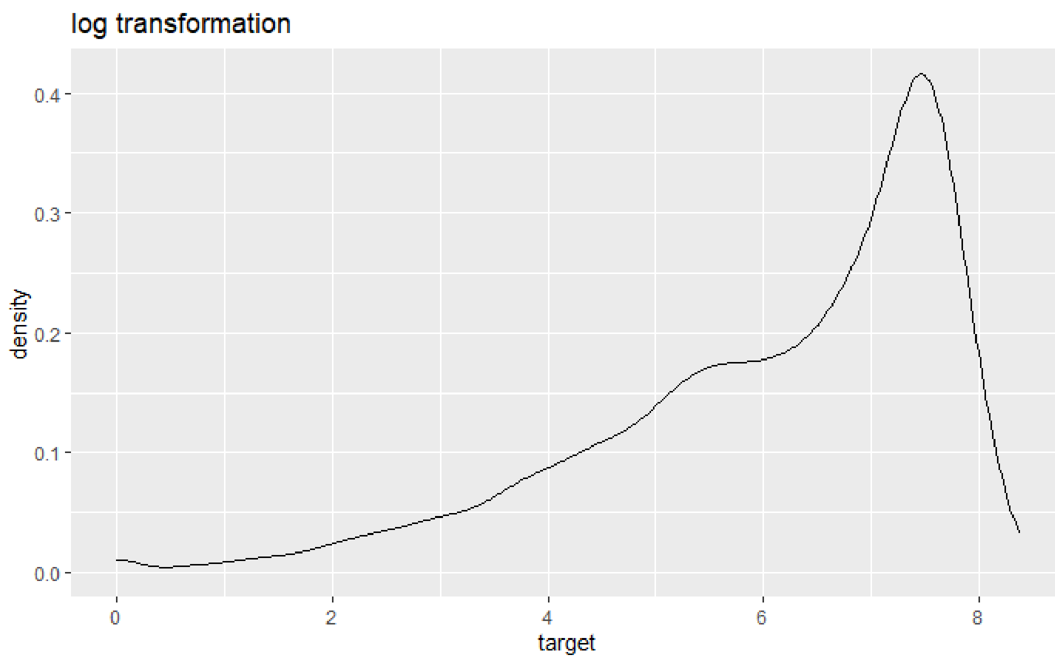
To illustrate these points, I trained a decision tree with a single predictor and a single split using our untransformed **pedestrians** variable, log(**pedestrians**), and square root(**pedestrians**). As discussed in the second part of this task, these represent right-skewed, left-skewed, and less-skewed distributions respectively. The resulting decision trees are shown below.

```
[1] "Right-Skewed Tree:"
n= 11373

node), split, n, deviance, yval
      * denotes terminal node

1) root 11373 8988081000  962.5301
  2) precipitation>=0.00515 1166  352529400  457.0369 *
  3) precipitation< 0.00515 10207 8303576000 1020.2750 *
[1] "Mean Prediction: 962.530115185087"


[1] "Left-Skewed Tree: "
n= 11373

node), split, n, deviance, yval
      * denotes terminal node

1) root 11373 30651.720 6.075959
  2) precipitation>=0.01265 798  2034.937 4.954004 *
  3) precipitation< 0.01265 10575 27536.470 6.160623 *
[1] "Mean Prediction: 450.429282855746"


[1] "Less-Skewed Tree: "
n= 11373

node), split, n, deviance, yval
      * denotes terminal node

1) root 11373 2740341.0 26.86222
  2) precipitation>=0.00715 1014  138994.6 16.94257 *
  3) precipitation< 0.00715 10359 2491802.0 27.83321 *
[1] "Mean Prediction: 731.210592771125"
```

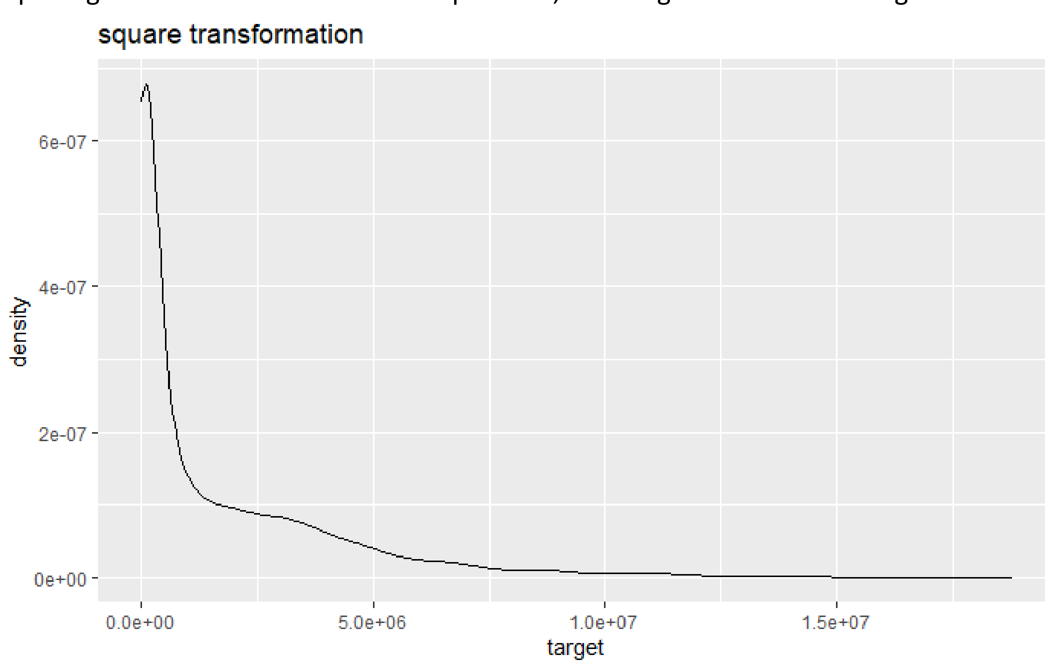
As expected, each tree split at a different precipitation value (highlighted in yellow), resulting in different numbers of observations in the leaves (highlighted in green). Also, we can see from the mean predicted values (highlighted in cyan) that the right-skewed version had the highest mean prediction, followed by the less-skewed version, and then the left-skewed version.

Picking a transformation

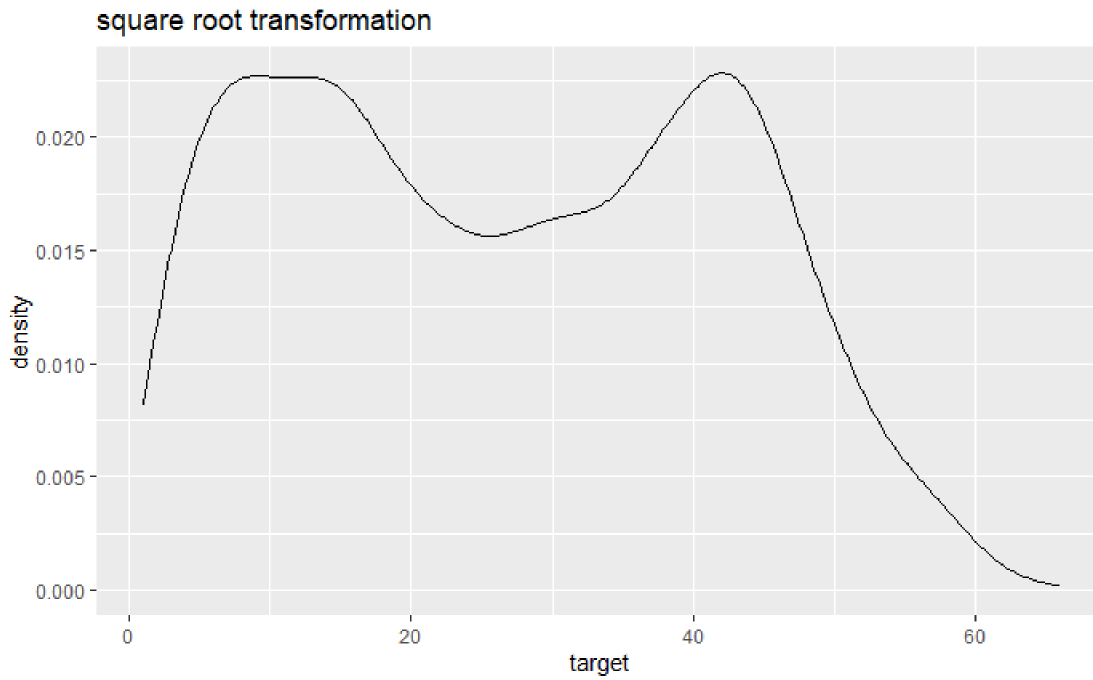
There are no zeros, so the log transformation is a viable option. Unfortunately, the transformation is too severe and causes the distribution to be left skewed to a similar degree.



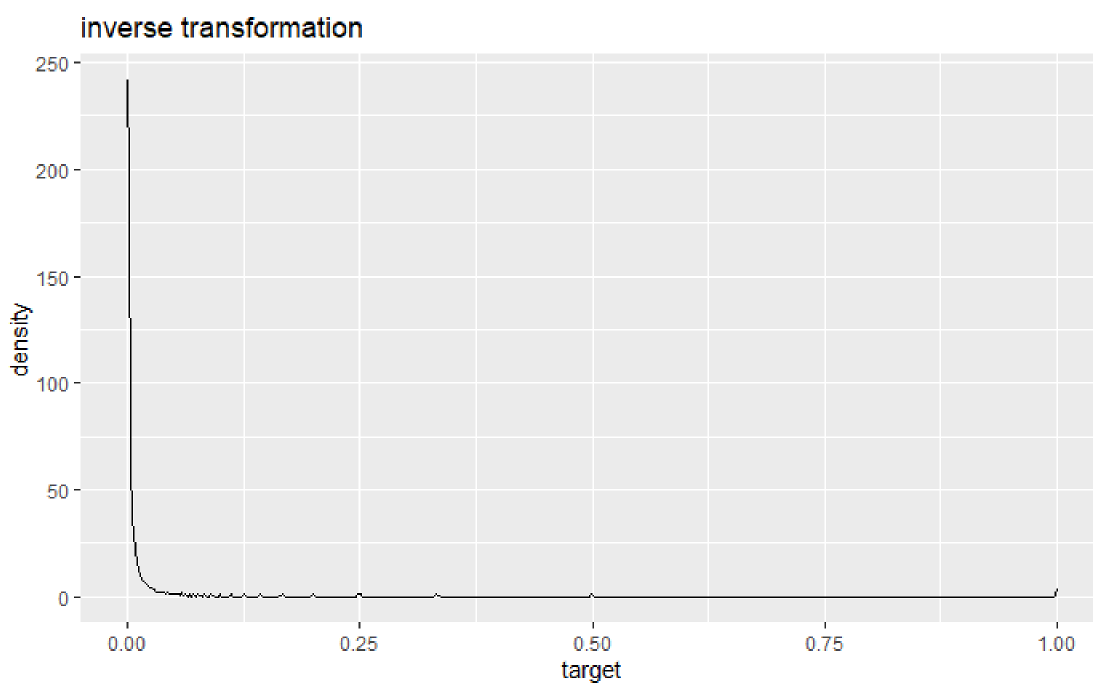
Squaring the variable exacerbates the problem, resulting in an even more right skewed.



Applying the square root will decrease the skewness.



Inverse will not work. The distribution is extremely right skewed, increasing the problem we are trying to solve.

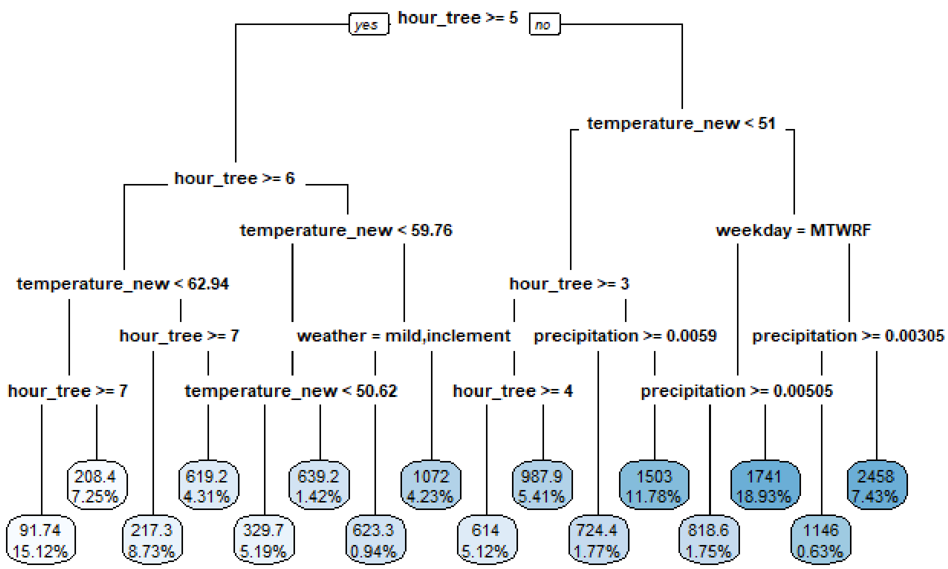


I recommend using the square root transformation. Of the transformations given, it was the only transformation that decreased the skewness of the target variable.

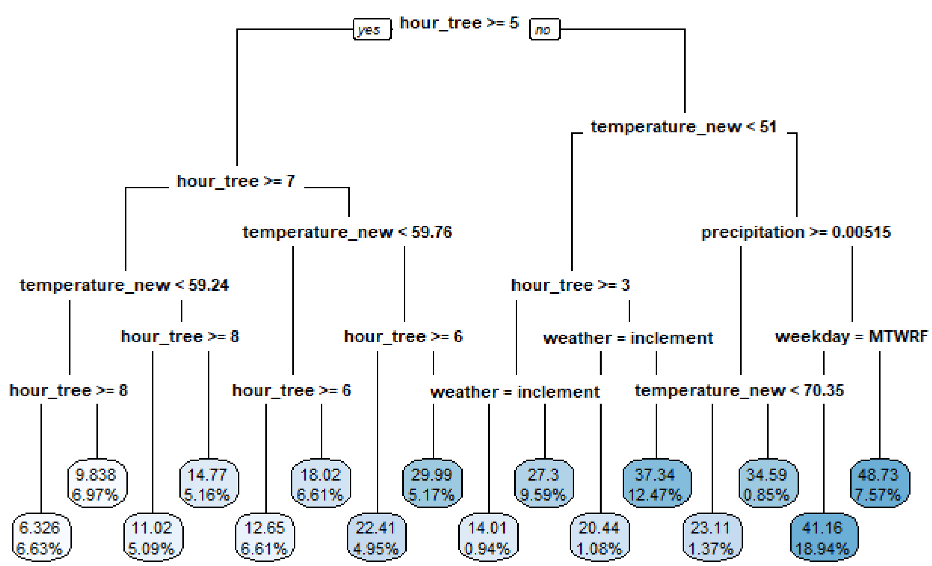
Task 5 – Build two trees (8 points)

Some candidates discovered that, after choosing hour_2 (the factor variable) for hour_tree in Task 3, the code provided in Task 5 did not work, and the section of code indicated not to change the code. The examination committee decided that the fairest solution was not to grade this task. All candidates received a score of 0 and the total points for the examination was reduced to 92..

Tree 1 (untransformed pedestrians):



Tree 2 (sqrt pedestrians):



To get the predicted pedestrians for each scenario, we first need to apply any transformations we used on the new data. For example, instead of hour = 6 in scenario A, we must use hour_tree = abs(14 – 6) = 8. After all of the transformations are applied, we can either use the predict function or manually follow the decision tree plot above. The table below shows the predicted number of pedestrians for each tree and scenario obtained using the predict function.

Scenario <fctr>	**Tree1_Predicted_Pedestrians** <dbl>	**Tree2_Predicted_Pedestrians** <dbl>
A	92	40
B	1146	1196

My assistant is overlooking the fact that the dependent variables are using different scales. In tree2, the target variable has been transformed using the square root function. For this reason, the RMSE's are not comparable. Furthermore, attempts to put the predictions on the same scale and then measure the RMSE do not yield valuable insights. When comparing the RMSE's using the original scale of the pedestrian variable, tree1 appears to be better, but if you compare the RMSEs using the square root scale, tree2 appears to be better.

Model	**RMSE (original scale)**	**RMSE (square root scale)**
tree 1	487.45	8.1311
tree 2	497.58	7.9857

We shouldn't be surprised by the results above. RMSE squares the errors before averaging, causing larger errors to have more weight than smaller errors, which is what we were trying to avoid when considering a transformation of the target variable. Note that the sum of squared errors used to fit the regression tree is a component of RMSE. Tree1 was built to minimize RMSE on the original scale, and tree2 was built to minimize RMSE on the square root scale. Hence, comparing the two trees based on RMSE on either scale is not helpful. If it is more important that the client can accurately predict high pedestrian counts than low pedestrian counts, then using RMSE on the original scale is appropriate.

Task 6 – Consider a random forest (3 points)

Almost all candidates were able to provide two quality considerations for this task.

Two other considerations when deciding whether to use a random forest:

1. Random forests don't lend themselves to illuminating the process that produced the data. Because a random forest is an ensemble of many decision trees that each might split the input variables differently, it will be difficult to use them to identify and interpret factors that relate to higher or lower counts of pedestrians, which is our goal.
2. Compared to a single decision tree, random forests are less prone to overfitting, so they are likely to generalize better to new data.

Task 7 – Fit a generalized linear model (8 points)

Many candidates mentioned that the Poisson distribution is commonly used for count data, but failed to provide any further justification for its use. Higher-scoring candidates were able to identify the nature of the target variable: discrete, positive, and right-skewed. They identified that the Poisson distribution is suitable for this type of data, and they suggested alternatives such as the gamma, inverse Gaussian, or lognormal distributions that could also fit positive, right-skewed data. Most candidates were able to successfully fit two separate models (Poisson + an alternative) and provide a recommendation based on a quantitative (e.g. lower RMSE) or a qualitative justification.

Our target variable distribution is non-negative and right skewed, so we should seek a similarly shaped probability distribution.

Poisson

Poisson makes the most sense because it has a similar distribution (non-negative and right-skewed) and is discrete, like our response variable. The Poisson distribution is the most common choice for count data.

Two other options

The gamma and inverse Gaussian distributions are also non-negative and right-skewed. Neither is discrete, but we can still model a discrete relationship with a continuous distribution. The inverse Gaussian distribution is more severely right skewed than the gamma distribution.

The GLM with Poisson and gamma distributions and log links achieve the following performance.

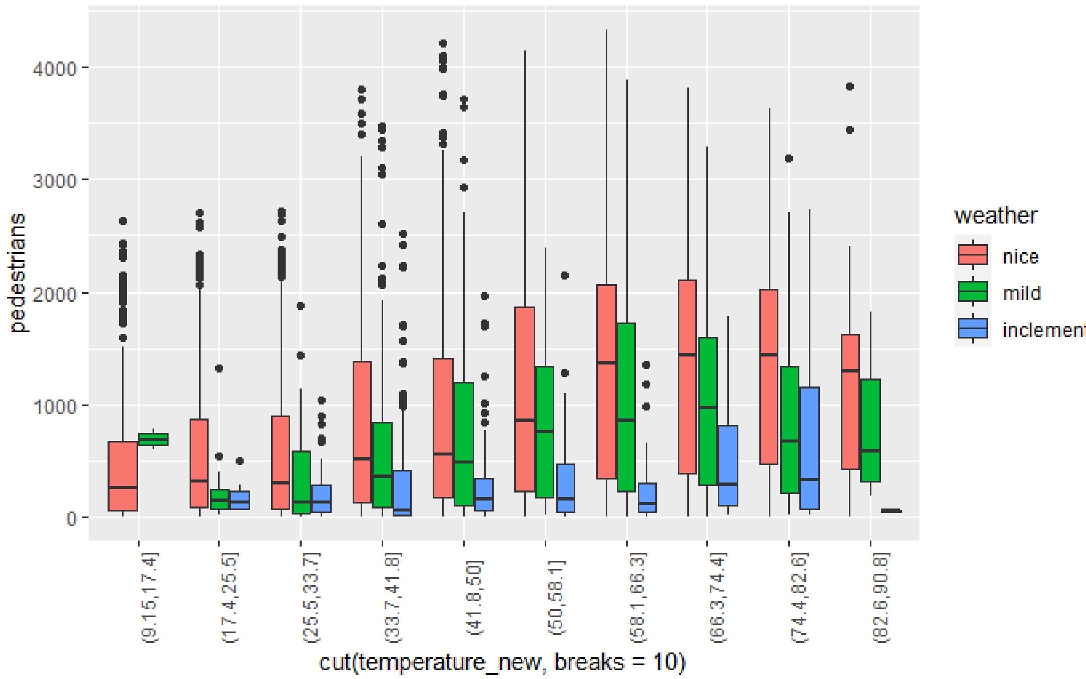
Model	Train RMSE	Test RMSE
GLM with Poisson and log link	561.37	563.33
GLM with gamma and log link	997.31	948.78

I recommend using the GLM with Poisson distribution because it had a lower RMSE and was the natural choice for count data, like our target variable.

Task 8 – Consider an interaction (6 points)

Some candidates struggled with this question, confusing association and interaction. To identify interaction: candidates should consider whether the predictor's effect on the target depends on the value of another predictor. Lower-scoring candidates frequently suggested an interaction between weather (e.g. clear, rainy) and precipitation. These variables are correlated; the "rainy" level of the weather variable co-occurs with precipitation. When interpreting the coefficients of the interaction, it was useful to discuss the main effects of the variables involved in the interaction, but it was not required to earn full points. The binned boxplot used below helps in visualizing the interaction, but the provided scatterplots could also be used to earn full points.

The following plot shows the relationship between **temperature_new** bins and the count of pedestrians grouped by the weather variable. I chose to bin the **temperature_new** variable for illustration purposes only, the actual variable is continuous. The rate of increase in pedestrian traffic as **temperature_new** increases is clearly different for inclement weather than other weather conditions. Most of the inclement weather records involved rain. In rainy conditions, there may be a temperature threshold below which temperature differences do not matter because the pedestrian would still end up cold and wet. Above that threshold, there may be some pedestrians that are willing to get wet since they won't be cold too. The difference in pedestrians between nice and mild also becomes more differentiated in warmer weather.



The GLM with the interaction had the coefficient estimates shown in the table below.

```
Coefficients:
                               Estimate Std. Error  z value Pr(>|z|)    
(Intercept)                   7.051e+00  1.447e-03  4873.14   <2e-16 ***
weathermild                  -5.612e-01  4.809e-03  -116.70   <2e-16 ***
weatherinclement             -1.579e+00  9.296e-03  -169.88   <2e-16 ***
precipitation                -5.751e+00  4.885e-02  -117.73   <2e-16 ***
weekdaySaturday               3.964e-01  9.307e-04   425.91   <2e-16 ***
weekdaySunday                 1.911e-01  1.025e-03   186.53   <2e-16 ***
hour_glm                     -2.924e-01  1.668e-04 -1752.68   <2e-16 ***
temperature_new               1.384e-02  2.176e-05   635.92   <2e-16 ***
weathermild:temperature_new   5.964e-03  8.210e-05    72.64   <2e-16 ***
weatherinclement:temperature_new  1.522e-02  1.566e-04    97.21   <2e-16 ***
```

Interaction coefficient interpretation

- When **weather** = "nice" there is no additional effect from the interaction with **temperature_new** because "nice" is the base level of the **weather** variable.
- When **weather** = "mild", the predicted count of **pedestrians** is multiplied by $e^{0.005964}$ = 1.005982 for each unit increase of **temperature_new**, on top of the multipliers already introduced by the main effects. So the predicted count increases by about 0.6% for each additional unit increase in **temperature_new**. The predicted count increases at a faster rate per unit increase in

temperature_new under "mild" than when the weather is "nice". This effect looks different than what is seen in the plot above because other factors also affect the comparison of "nice" and "mild" at different temperatures.

- When **weather** = "inclement", the predicted count of **pedestrians** is multiplied by $e^{0.01522}$= 1.015336 for each unit increase of **temperature_new**, on top of the multipliers already introduced by the main effects. The predicted count increases at the fastest rate per unit increase in **temperature_new** for inclement weather.

Task 9 – Select features (9 points)

Most candidates were able to successfully describe forward and backward selection, but many struggled to interpret the results of the first step of the StepAIC procedure. Stronger candidates went beyond discussing the mechanics of the two methods and also addressed tendencies of the approaches to result in simpler models, more complicated models, or stop at local minima.

Because we will be using BIC to select features, only BIC is used in the discussion about forward and backward selection, even though a different criterion could be used.

Forward selection begins with a GLM that only has the intercept term. At each step, each possible predictor variable in the dataset that is not already included in the model is then *added*, one at a time and then substituted by the next, and the predictor whose *addition* results in the lowest BIC is *added* to the model. The process repeats until the BIC cannot be reduced by *adding* predictors to the model.

Backward selection begins with a "full" model that uses all of the predictors. At each step, each possible predictor variable still included in the model is *removed*, one at a time and then substituted by the next, and the predictor whose *removal* results in the lowest BIC is *removed* from the model. The process repeats until the BIC cannot be reduced by *removing* predictors from the model.

Differences between forward and backward selection:

- Forward selection starts with an "empty" model, while backward selection starts with a "full" model.
- In forward selection, predictors are added to the model at each step, but in backward selection, variables are removed at each step.
- Forward selection is more likely to get stuck in a local minimum BIC model that is too simple, whereas backward selection is more likely to get stuck in a local minimum BIC model that is too complex.

I recommend using forward selection because it is more likely to result in a simpler model which will be easier to interpret for the client.

The following features, which was all of them, should be retained based on the output of the forward selection.

- hour_glm
- temperature_new
- weather

- weekday
- precipitation
- temperature_new:weather

The output from the first step of the stepAIC procedure is:

```
Start:  AIC=6877716
pedestrians ~ 1

                  Df Deviance     AIC
+ hour_glm         1  3297189 3360255
+ temperature_new  1  6245615 6308681
+ weather          2  6414172 6477246
+ precipitation    1  6540826 6603892
+ weekday          2  6605687 6668762
<none>                6814660 6877716

Step:  AIC=3360255
pedestrians ~ hour_glm
```

Even though the output says "AIC" repeatedly, these are BIC values. The initial model with no predictors and only an intercept term has a BIC = 6877716. In this step, a model using only an intercept and one of the predictors in the left-hand column was built, and each of those resulted in the BIC values in the right-hand column. The number of degrees of freedom added by a variable is indicated, with factor variables sometimes adding more than one degree of freedom because all levels of a factor variable are added at once. Because the model that added the **hour_glm** variable resulted in the lowest BIC value, that model was used to start the next iteration.

Task 10 – Recommend a model (8 points)

Candidates generally performed well on this question. Weaker candidates often failed to provide any qualitative rationale for their selection. Candidates were required to run the recommended model on the full dataset in preparation for the executive summary. Even though there was no requirement that candidates mention running their recommended model on the full dataset in preparation for the executive summary, evidence of performing this step needed to be present in the RMD file to earn full credit.

Recommended GLM from task 7, 8, and 9

The GLMs from task 8 and task 9 are the same. The RMSE results on the train and test sets of the models are shown in the table below.

Task Source	Model	Train RMSE	Test RMSE
7	GLM with Poisson and Log link	561.37	563.33
8	GLM with Poisson and Log link + Interaction	558.09	561.74
9	GLM with Poisson and Log link + Interaction based on Forward Selection (Same model as task 8)	558.09	561.74

The performance of the GLM with the interaction term is superior to the GLM without the interaction, but the presence of the interaction term makes the model more difficult to interpret.

The marginal improvement in predictive performance does not seem worth the tradeoff of a more difficult interpretation, so I recommend the GLM without the interaction term (from task 7).

Recommended Model between GLM Above and the Decision Tree

Model	Train RMSE	Test RMSE
GLM with Poisson and Log link	561.37	563.33
Decision Tree	476.81	487.45

- The client is more interested in understanding the factors that lead to more or less pedestrian traffic than they are with model performance. Both model types are highly interpretable, so they will be able to meet that need.
- The decision tree from task 5 outperforms the recommended GLM by a *sizable* margin.
- The business problem noted that insight into how weather affects pedestrian activity would be useful for considering new locations across the U.S. We know the data used to train and test the models is from New York City, and the decision tree may not perform well when applied to a new location because decision trees are very sensitive to changes in the data. The directional insights learned from the GLM might be more applicable to new localities than the split points identified by the decision tree.

Even though there are concerns about how the decision tree results will generalize to new areas, it still is an easy-to-interpret model that allows us to gain insights for the client, and the improved performance compared to the GLM suggests the insights we gain from the decision tree may be more likely to be of value. Therefore, I recommend using the decision tree over the GLM.

Task 11 – Validate the model selection (4 points)

Many candidates simply stated the RMSE of the final selected model using the holdout data and provided a brief comparison. Some candidates inappropriately trained a new model using the holdout data, then compared RMSE from that trained model to the RMSE when trained using the training data or using the full data set.

Model	Train RMSE	Test RMSE	Holdout RMSE
GLM with Poisson and Log link	561.37	563.33	597.20
GLM with Poisson and Log link + Interaction (same for task 8 and 9)	558.09	561.74	594.36
Decision Tree	476.81	487.45	476.43

On the holdout set, the final recommended decision tree (highlighted) retained its relatively high performance. The GLMs both had larger drops in performance on the holdout set compared to the train set than the decision tree. Given this, the fact that the decision tree continued to perform well on the holdout set adds confidence to the choice of the decision tree as our final model.

Task 12 – Executive summary (20 points)

Candidates were expected to write their summary using language appropriate for the audience. The use of technical terms should be avoided if possible and accompanied by clear definitions otherwise. Unexplained performance metrics, describing results in terms of square root pedestrians rather than actual pedestrians, and describing the detail in extreme detail using terms like "integer" and "factor" are a few examples where candidates failed to meet expectations.

The focus of the summary should be on key details of interest to the client. Most candidates did well describing the data and any adjustments to the data, but they rarely pointed out important limitations of the data source such as it being collected in only in one place. Candidates did well when describing key insights for the client, but had difficulty describing the model development, performance, and recommending next steps.

To: A National Retail Firm

From: Actuarial Analyst

You have asked us to identify and interpret factors that relate to higher or lower levels of pedestrian activity. Insights about drivers of pedestrian activity are expected to be useful when considering new locations across the U.S. for stores and determining the optimal hours for having stores open since your sales are proportional to pedestrian activity. To assist you with this problem, we have built a model to predict pedestrian traffic levels that we believe provides such insights.

The data studied is from NYC Open Data and were collected in 2017-2019. The 11,373 records contain information about the number of pedestrians observed during each hour, the time of day (in one-hour segments), the day of the week, and various weather information. Prior to our work, the data had already been cleaned. Because the data was collected in New York City, insights gained from the data may not translate well to other places where climate patterns, public transportation, and pedestrian-friendly city planning may differ.

Preparing for Modeling

To begin, we explored whether or not the variables were likely to predict the number of pedestrians based on data exploration and domain knowledge. We determined that the time of day variable was most likely to be important, and the temperature variable was least likely to be important because it duplicated information contained in the forecasted daily average temperature.

Next, we transformed several variables to make them more suitable for modeling. Some variables simply had too many categories, and that would have led to an overly complex model that would not be a good predictor of future pedestrian activity. Additionally, many of the variables in the dataset were related to each other in ways that would make it difficult for the models to separate their impacts. After analyzing the data, we made the following changes that impacted our final model:

- Rather than considering each of the seven days of the week on their own, we grouped Monday through Friday since they are workdays, and considered those separately from each of Saturday and Sunday.

- We regrouped the weather variable into 3 categories: “nice” (clear or partly-cloudy), “mild” (cloudy, windy, or fog), and “inclement” (rain, snow, or sleet).
- We changed the time of day variable to measure the number of hours away from 2 p.m.
- We decided to use the forecasted daily average temperature instead of the hourly temperature measure.

Model Selection

Three types of models were considered for this project, decision trees, generalized linear models (GLMs), and random forests. Decision trees use a series of if-else statements to make a prediction. GLMs use a mathematical formula to make a prediction. Random forests combine the predictions of many decision trees built with random sampling to make a prediction. We concluded that a random forest was not a good fit for the project, as they were not likely to yield useful insights about the drivers of pedestrian traffic due to being difficult to interpret. Several GLMs and decision trees were trained and compared.

Our final model is a decision tree. To select this model, we considered (1) the ability to gain useful insights from the model, (2) the model performance as measured by root mean squared error (RMSE) – a commonly used metric for measuring model accuracy, and (3) the ability to apply those insights to new localities. The tree model is easy to interpret, so we will be able to extract meaningful insights. It performed significantly better than the other models as its predictions had smaller errors on average. The decision tree was not as good as the GLM for applying the useful insights to new localities, but the performance of the GLM was worse enough that we did not think the insights from it would be as useful in any localities, so the decision tree was selected.

Expected Performance

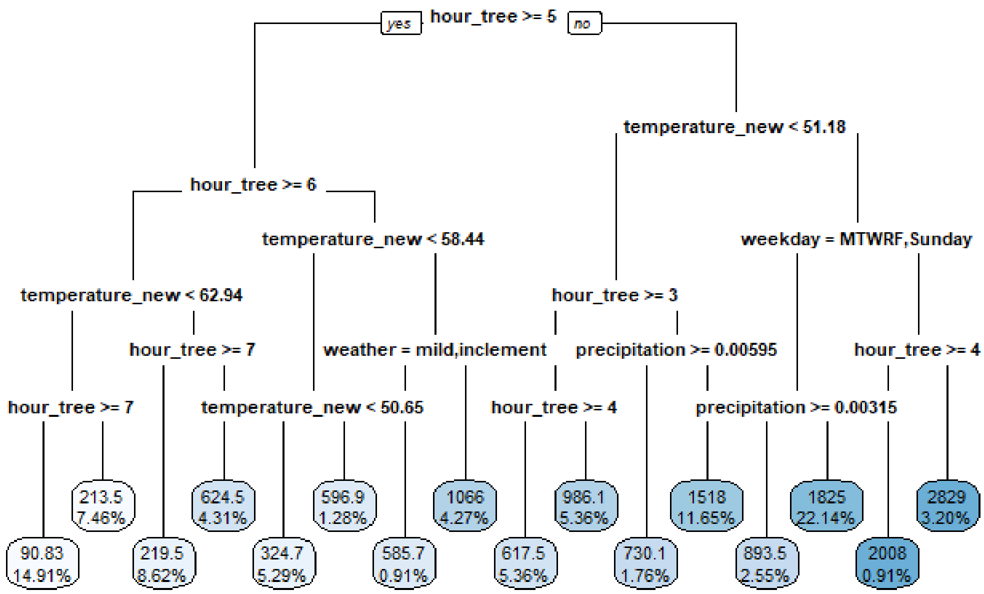
Before we discuss the insights learned from the model, it is important to verify that the model can reasonably predict pedestrian traffic. We estimated the RMSE performance you should expect when applying the model by measuring its effectiveness on a set of 1,136 unseen records (holdout data) that was not used to train or select the model. Our selected model RMSE was 476. To illustrate the value of the model, we can compare that performance to simply predicting that the average number of pedestrians would be present each hour of each day. On average there were 960 pedestrians for each record in the data we used to train our models. Simply predicting that 960 pedestrians would be present at any given time would have a RMSE of 907. A lower RMSE is preferred, so our model is adding value.

The Model

The following 5 factors were important for determining the number of pedestrians (listed in order of importance):

- The number of hours difference between the current time and 2 p.m.
- The average temperature forecast
- The hourly rate of precipitation in inches
- Whether it was Saturday, or a different day of the week
- Weather conditions as categorized in the second bullet under “Preparing for Modeling” above

The decision tree is displayed below:



To use the tree to make a prediction, follow a branch of the tree downward based on the answers to the question at the top of each split ("no" always goes right, while "yes" always goes left), eventually arriving at one of the sixteen possible predicted values at the bottom of the tree.

As an example, consider the case where it is Saturday at 11 a.m., raining 0.01 inches per hour on a day where the average temperature forecast was 60 degrees Fahrenheit, and you want to predict the number of pedestrians. Begin at the top of the tree and answer a series of questions to follow the tree downward.

Question 1: Is it at least 5 hours from 2 p.m. in either direction? Because the answer is no, you would head down the right branch to the next question.

Question 2: Is the average temperature forecast less than 51.18 degrees Fahrenheit? The answer is no, so again you would head down the right side of the tree to the next question.

Question 3: Is it Monday through Friday or Sunday? It is not, so you would head further down the right side to the final question.

Question 4: Is it at least 4 hours from 2 p.m.? Because it is not, you would head down the right side to the predicted pedestrian count, which is the top number in the circle at the bottom, 2,829 pedestrians.

Insights

- The time of day is important. In fact, whether it was between 9 a.m. and 7 p.m. was the most important driver of pedestrian traffic, with more between those times.
- Colder average temperature forecasts were associated with fewer pedestrians, but the temperature forecasts leading to more or less pedestrians is dependent on the time of day.

- The weather conditions are only important at some combinations of time of day and average temperature forecast.

These results make intuitive sense. Our decision tree found that there is less pedestrian activity when in early morning, late evening, or when most are sleeping at night. Additionally, more pedestrians are present in forecast temperatures and weather conditions that are more comfortable for walking.

Next Step

One weakness of our final model is the fact that decision trees are sensitive to small changes in the data. Because our model was trained using data collected from just New York City, caution should be used when applying the existing model to new localities. We recommend seeking additional data from other localities to improve the model.