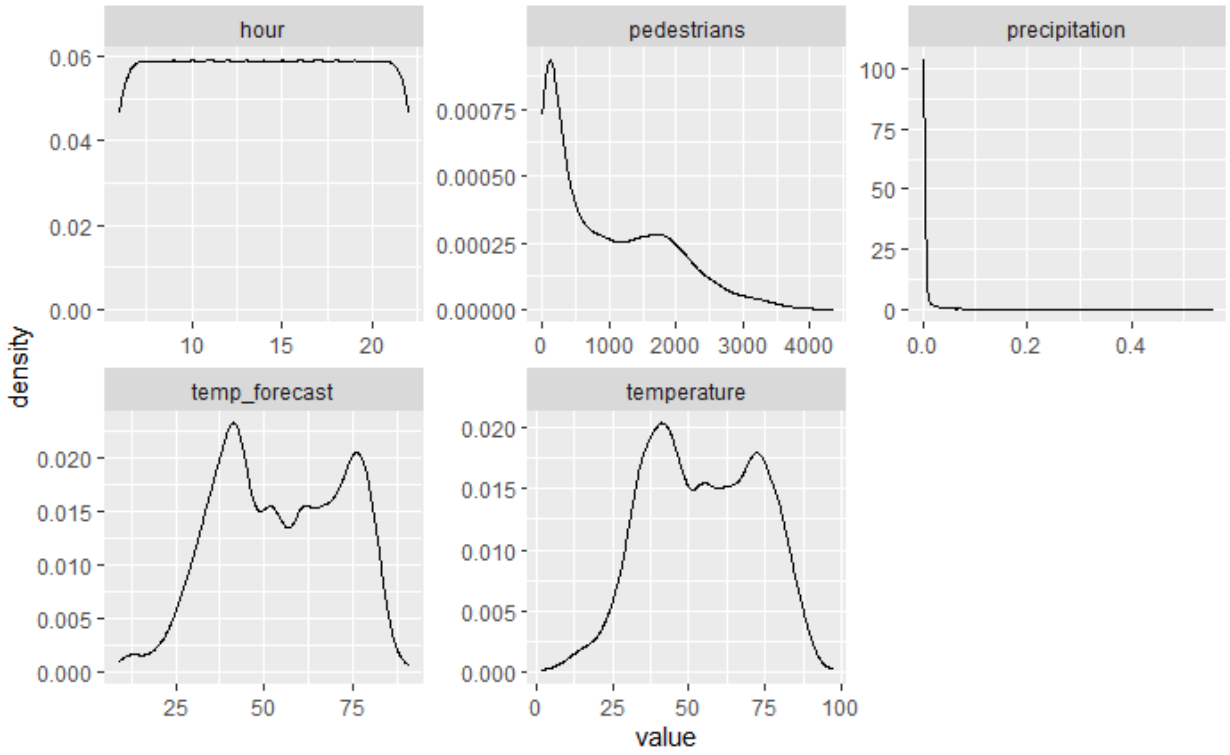
Practice Exam – Pedestrian Activity (SOA PA 12/8/20)

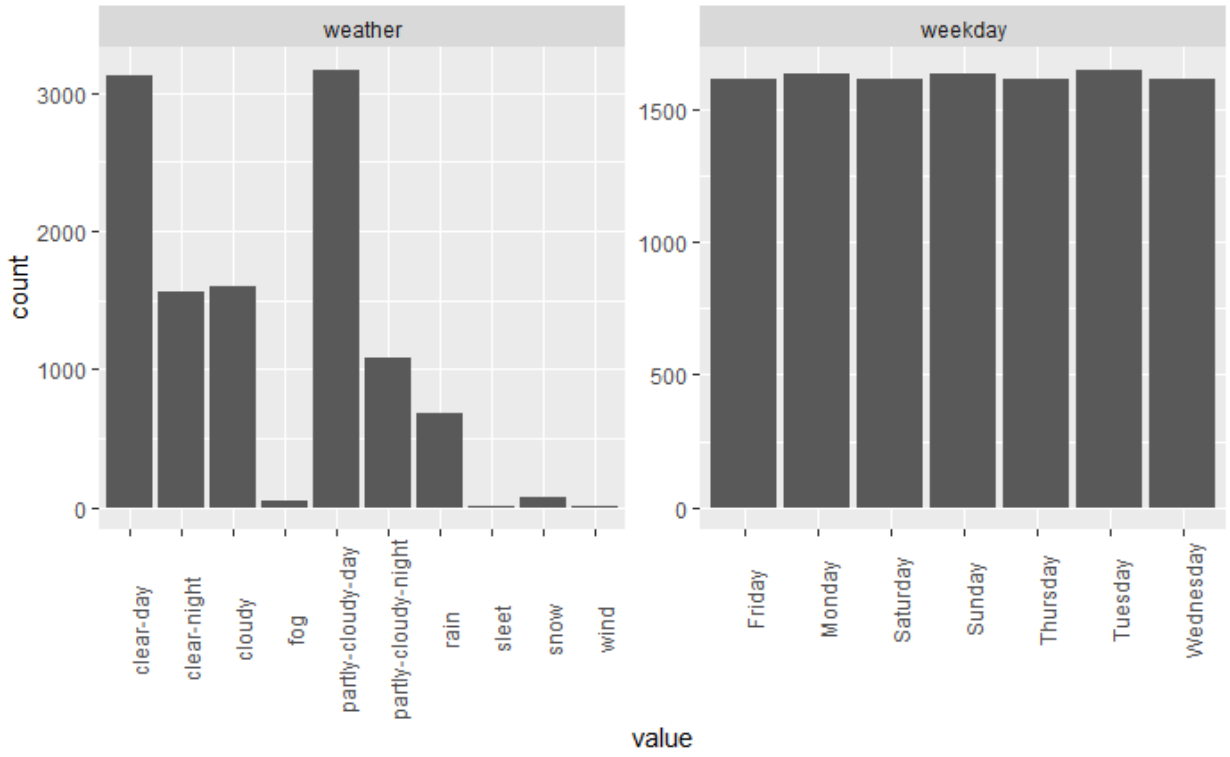
Instructions to Candidates: Please remember to avoid using your own name within this document or when naming your file. There is no limit on page count.

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)



These histograms are not informative of which will be predictive of **pedestrians** apart from showing the spread of values. If a variable had many zeros, or missing values, then we would be able to see this from the histogram. In this case, we do not make any conclusions.



This graph shows the counts and not **pedestrians**. Notice that weather has only a few values for rain, sleet, snow, wind, and fog and so these factor levels are of questionable credibility. There are the same number of records on all days of the week.

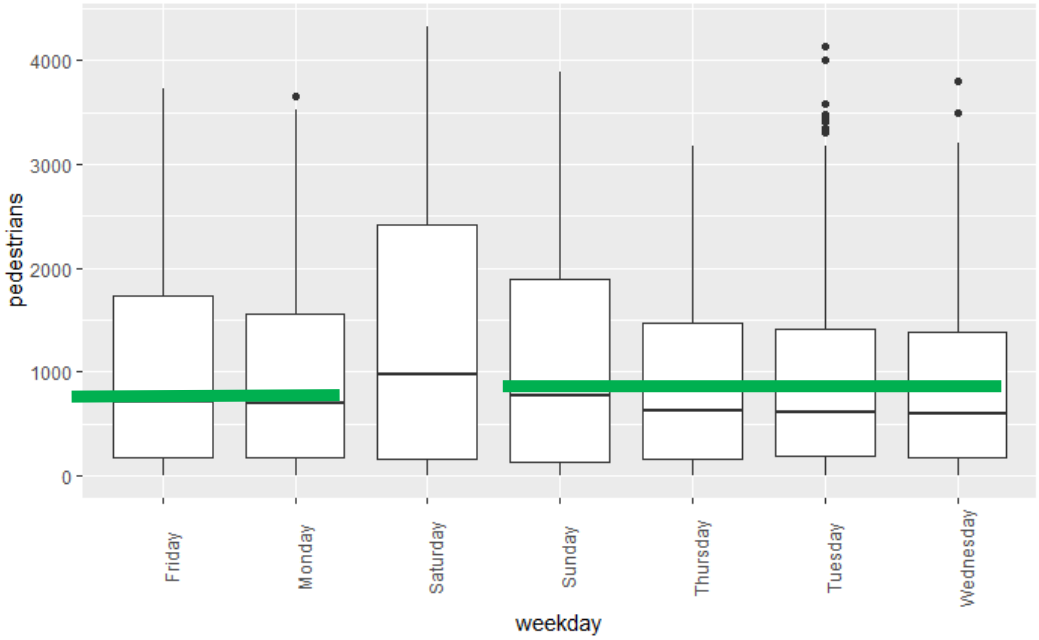
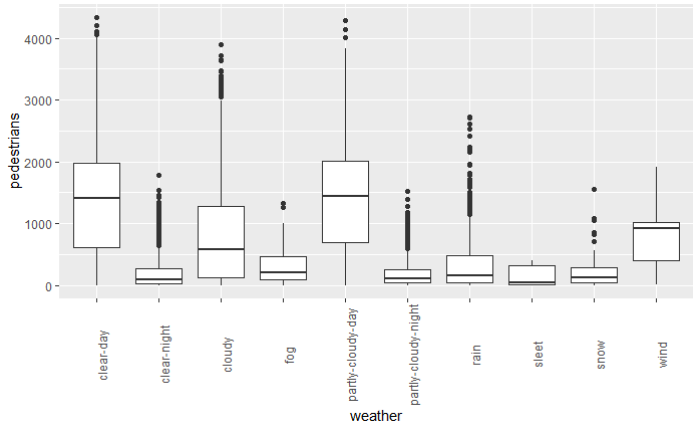
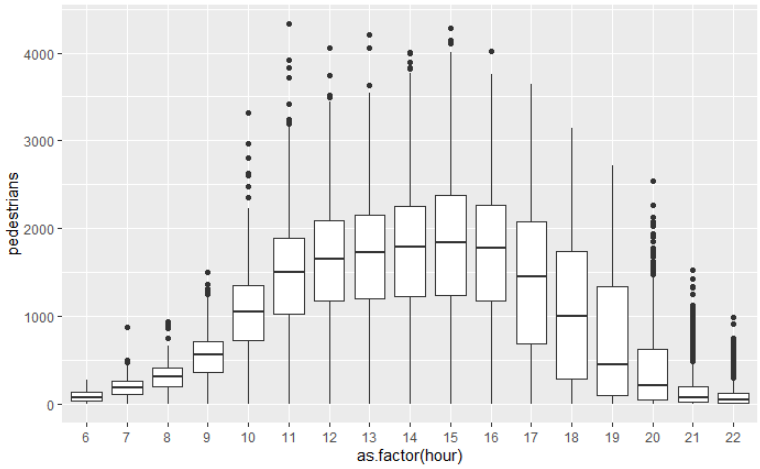


Figure 1 Green line shows overall trend

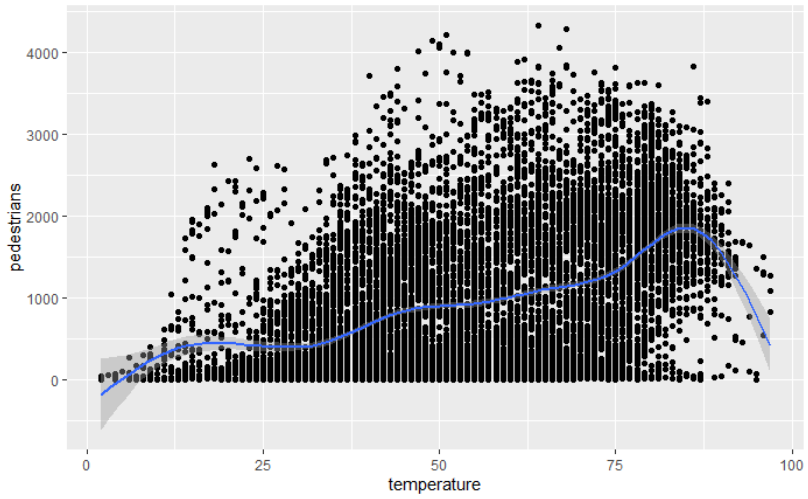
The green line in the boxplots of **weekday** shows that the median is about 800 pedestrians for all days of the week except for Saturday. Saturdays have about 1000. This implies that **weekday** is not predictive, however, we need to consider interaction terms with other variables before being able to make this conclusion.



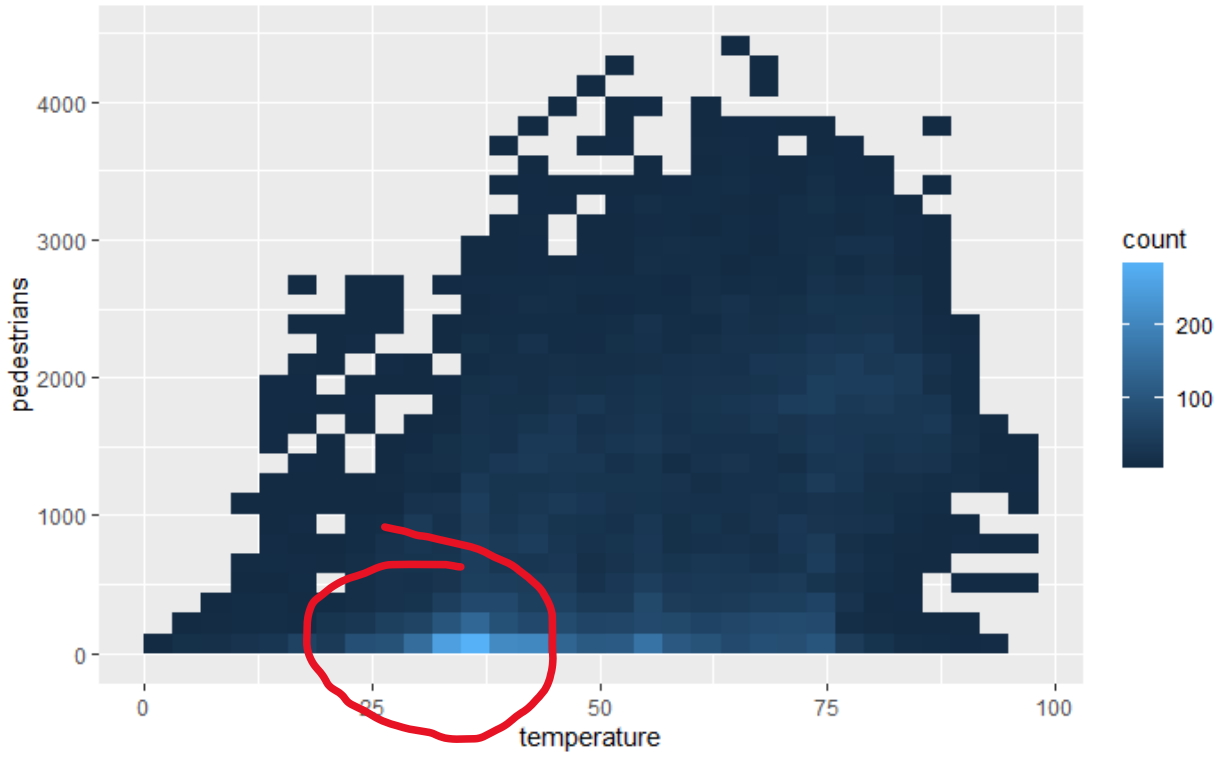
Weather is not reliable because there are many factor levels which have only a few records. There is also no clear distinction between what the levels represent. What does a "clear-day" look like? How does that compare to "Party Cloudy"? These are questions which do not have a definite answer. The data dictionary just says that this is the "hourly weather condition" but does not specify of how this is measured.



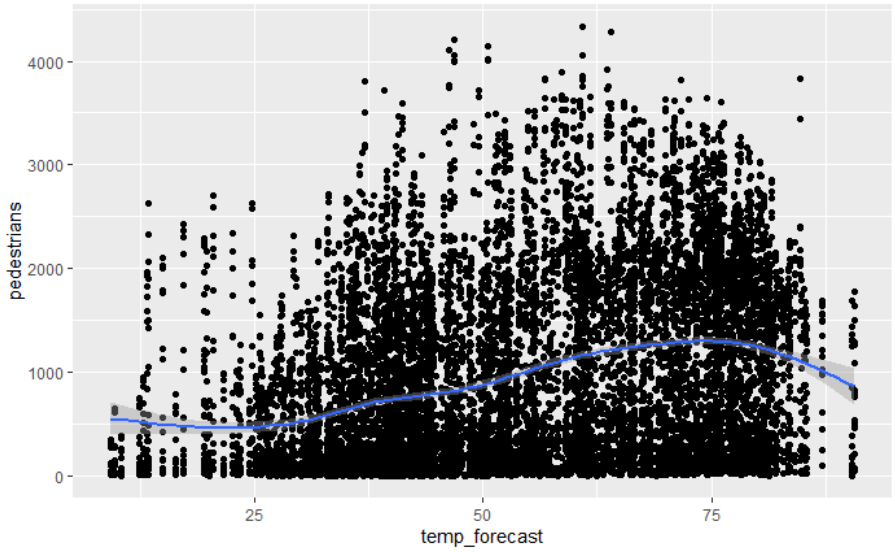
The time is measured by **hour** which records the number of pedestrians from 6 am through 10 pm. There is a peak in **hour** where most pedestrians go out during the middle of the day from 10 am through 5 pm. During the early morning and night, there are zero pedestrians whereas during the day there are between 1000 and 2000. This implies that **hour** is strongly predictive of **pedestrians.**



The **temperature** graph is difficult to interpret. The graph is not well specified. A choice of a dot plot here is poor because there are many points which lie on top of one another. The trend line, shown in blue, does help by showing the average **pedestrians**, however, a different type of graph would be better.



This bivariate histogram does a far better job of showing the distribution of observations across **pedestrians** on the y-axis and **temperature** along the x-axis. There is a concentration of points around the red circle that was not visible in the scatterplot. The fact that the graph is one solid color of dark blue shows that there are about the same number of observations scattered around randomly. This implies that temperature will not be predictive of **pedestrians** on its own.

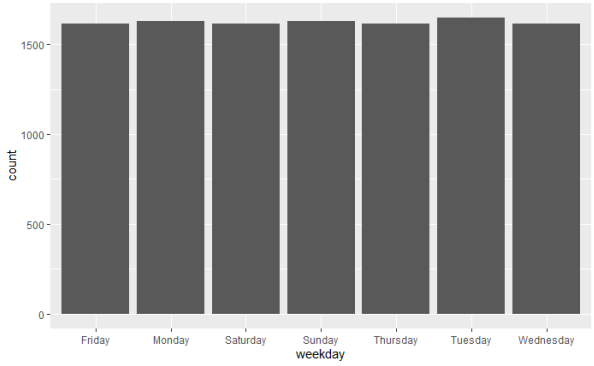


The data dictionary says that **temp_forecast** is the daily average temperature prediction whereas temperature is on an hourly level. This seems wrong because there should be 24 points in the temperature graph for every one point on the above graph. In any case, this variable does not seem to be prediction of **pedestrians** for the same reasons as **temperature.**

In summary, the **hour** is the most likely to be predictive and **temperature** is the least likely.

Task 2 – Reduce factor levels (7 points)

The weekday was regrouped to be either a Saturday, a Sunday, or a weekday. The base levels here are not important because all days have the same number of observations.



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

As was pointed out from the earlier box plots, the levels of fog, sleet, snow, and wind have few observations which implies that they should be grouped together into simpler groups.

clear-day	clear-night	cloudy	fog	partly-cloudy-day	partly-cloudy-night
3127	1565	1601	54	3169	1081

rain	sleet	snow	wind
679	11	77	9

There are several ways that this could be done. Simply, we could group into day or night. Alternatively, we could use good weather or poor weather. This binarization would be too simple for our problem of considering the time of day as well as the weather conditions and so we will use these groups instead.

Old level	**New Level**
Clear-day, cloudy, partly-cloudy-day	Nice day
Clear-night, partly-cloudy-night	Nice night
Fog, sleet, snow, wind	Bad weather
rain	rain

Most observations are with fair weather. Rain is put into its own group because the median of **pedestrians** is higher than fog, sleet, snow, or wind on the boxplot from Task 1. In general, it is bad practice to combine factor levels which have a different median value of the target because this throws out information.

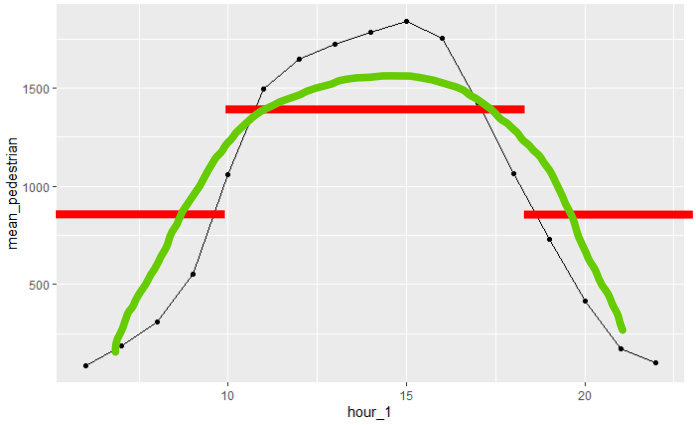
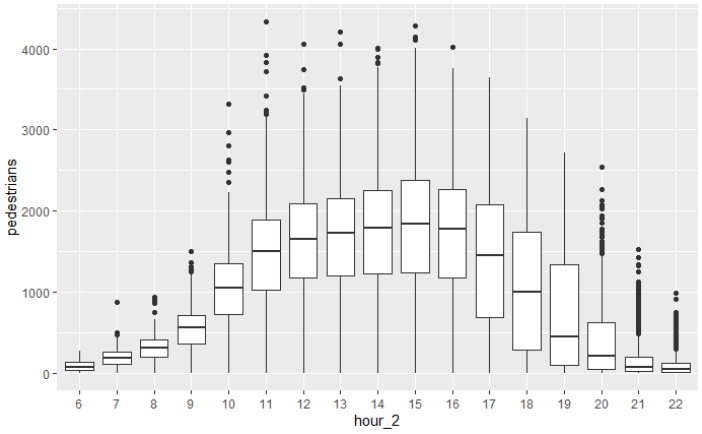
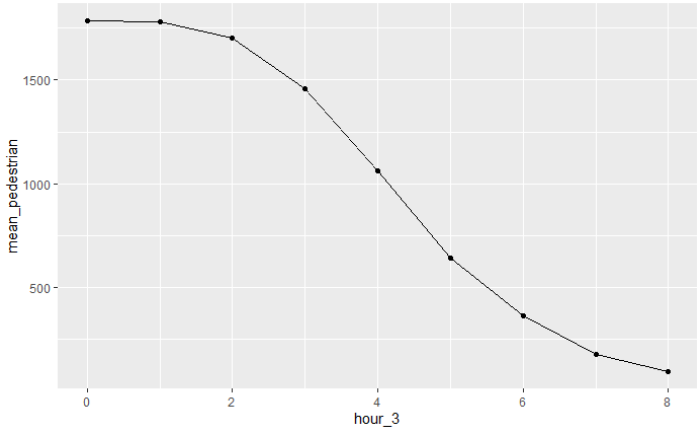


Figure 2 Red (Decision Tree), Green (GLM hour_1 squared)

Hour_1 is numeric and has a change from increasing to decreasing. This means that a decision tree could model it using different cut points. For example, the red lines show where a tree could make predictions of the mean of **pedestrians** for these regions.



Hour_2 is a factor which means that the tree could model the different levels as with hour_1, however, the result would be more complicated to interpret because there could not be a greater than or less than sign such as "10 <= hour_1 <= 25" but would need to have a rule like "hour_2 is in (10,11,12, 13, …, 25". For business purposes, this would be complicated to explain.



Hour_3 is numeric like **hour_1** and it is also approximately monotonic. This is recommended for the GLM because it only requires one coefficient to model. A GLM changes the mean of the response according to the value of the hour variable multiplied by the hour's coefficient. This results in a straight line in the predicted value which is closer to the above graph than **hour_1** is. Another option is to use a quadratic term **hour_1**^2 which would allow for a parabola as shown by the green line.

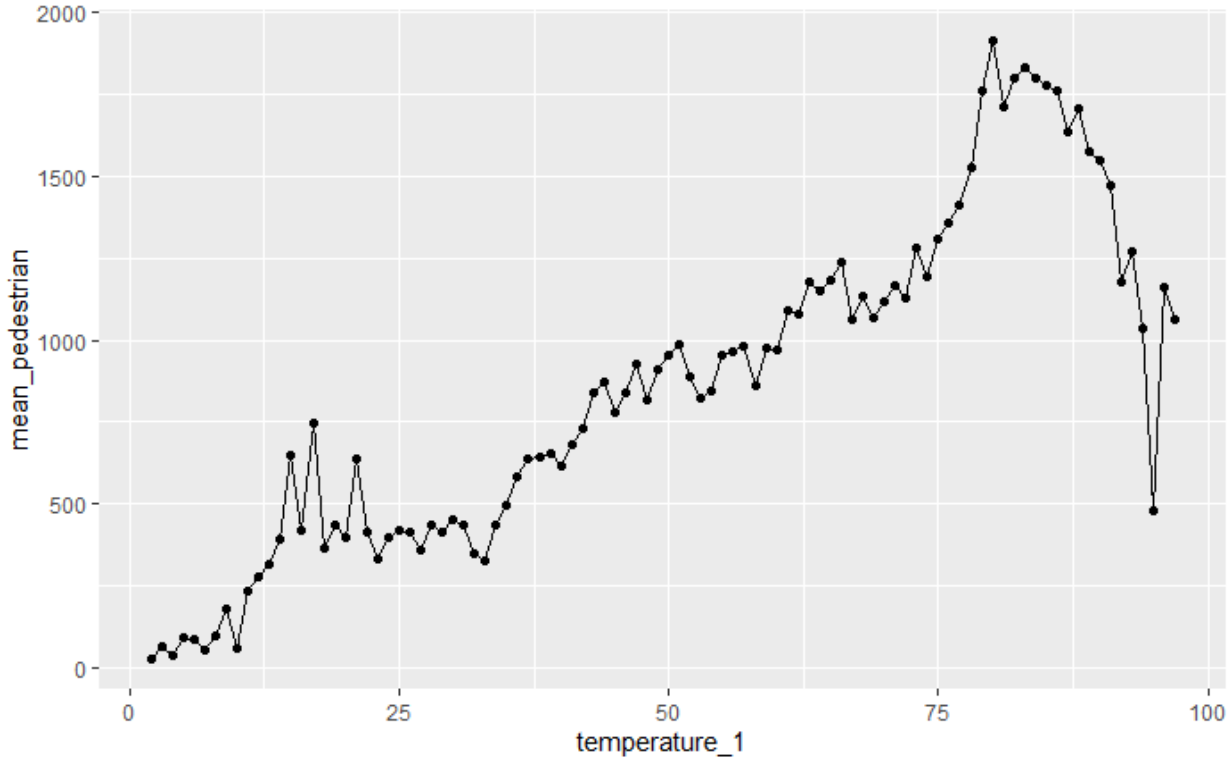
This message is in response to the situation you experienced while taking the December 2020 administration of the Predictive Analytics exam. After a review by SOA staff teams and volunteer leaders we determined that for candidates that chose hour_2 (the factor variable) for hour_tree in Task 3, the code provided in Task 5 did not work, and the section of code indicated not to change the code. The SOA has adjusted the grading of this version of the PA exam accordingly to take this issue into account and our final score will reflect these adjustments.

Thank you for bringing this issue to our attention and for your patience in waiting for a response.

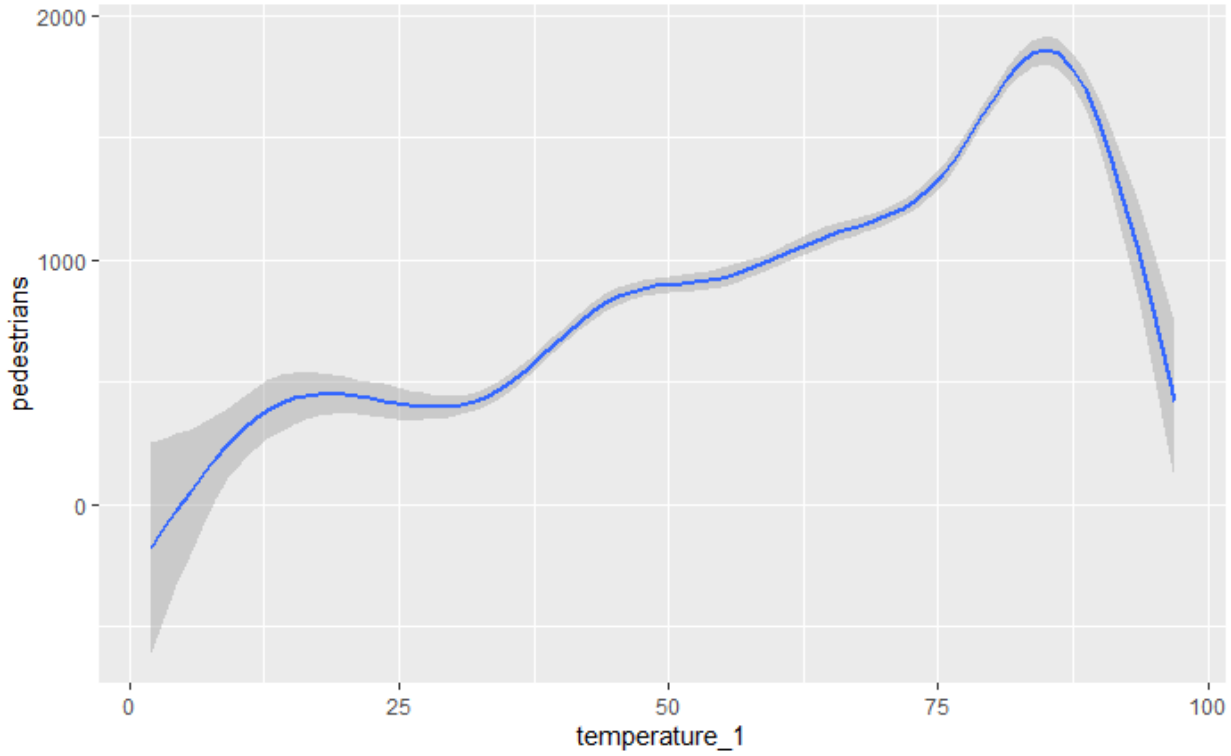
Kind regards,

PA Exam Staff

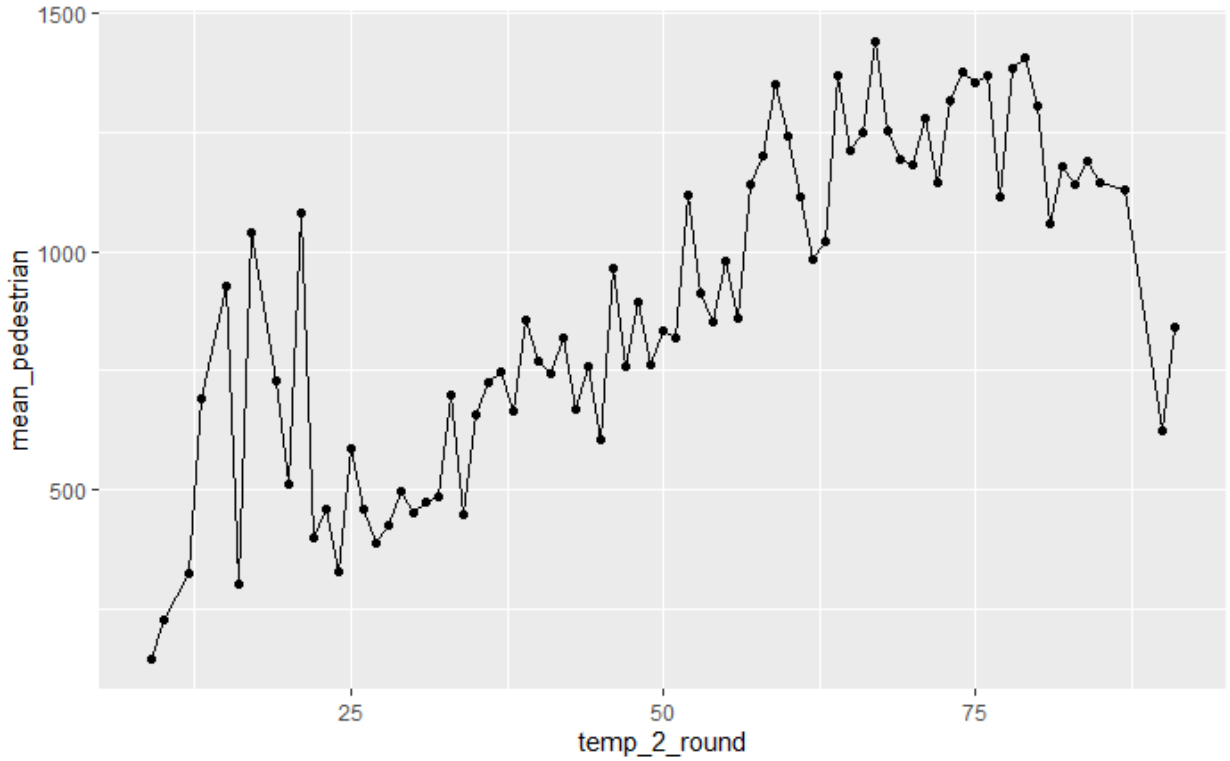
There was a technical error with the .Rmd file for this question. This email was sent to all candidates who took this exam in December of 2020. The best response was to comment this out and explain that it did not run correctly.



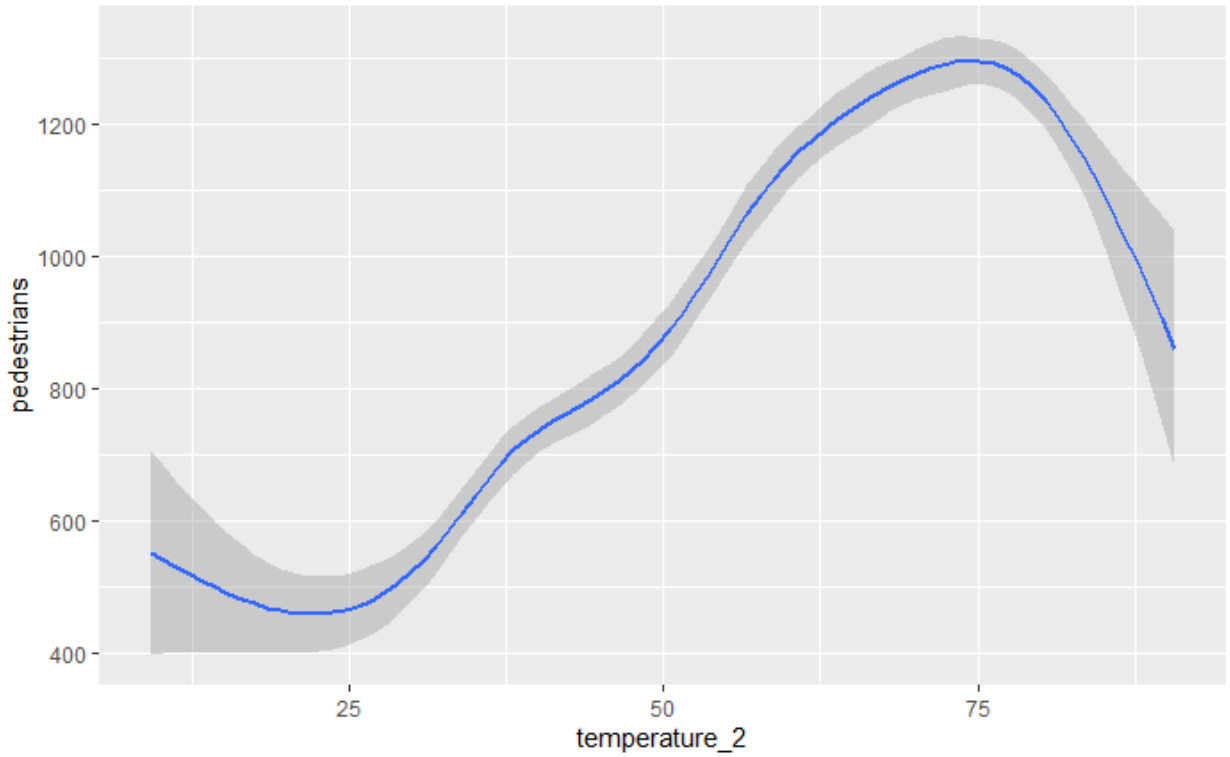
This graph shows the average **pedestrians** against the hourly temperature in degrees Fahrenheit.



This graph shows the uncertainty bands around the mean. A mean is a summary statistic which takes many values and combines them into one. The grey shaded areas get wider for lower temperatures because there is more variance here.



The **temp_2** variable is the predicted daily temperature, which should have fewer observations because there are 24 hours in one day, however, the data dictionary is miss labeling this. It appears that this is the hourly predicted temperature.

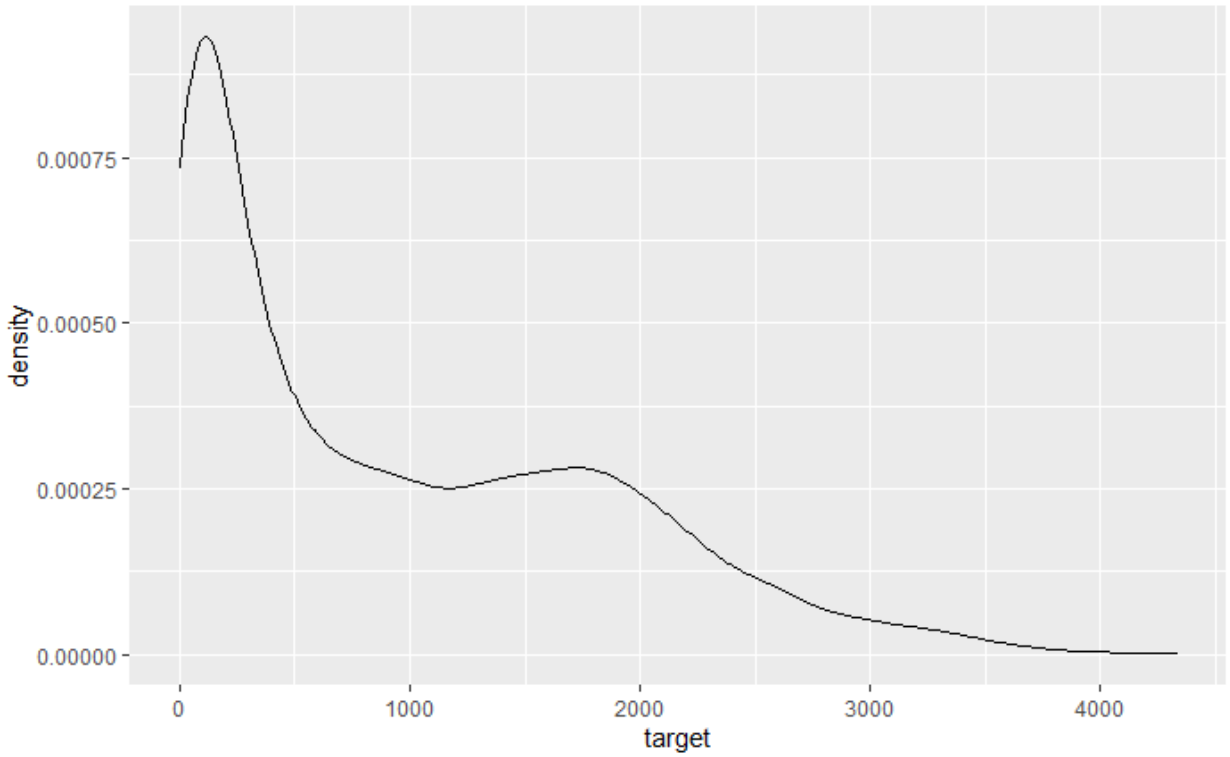


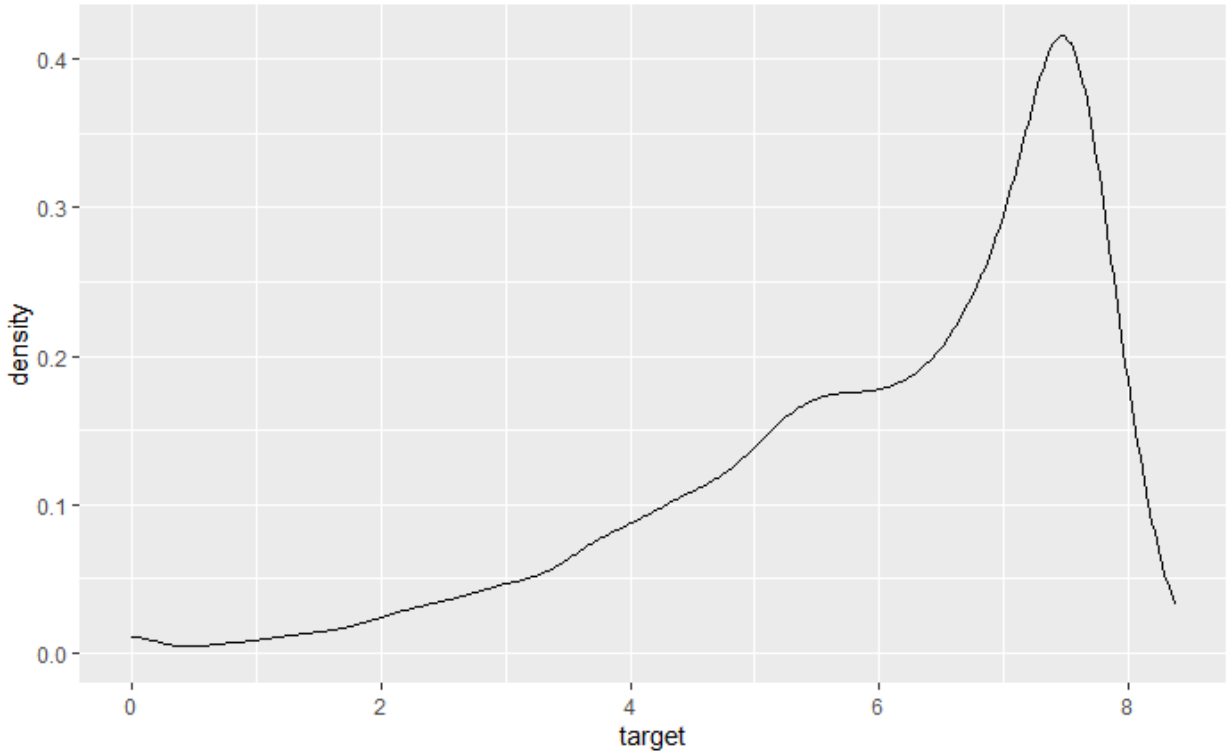
The confidence bands are wider on the predicted temp which means that the result is less reliable.

When deciding whether to use the actual temp or the predicted temp, we went back to the business problem which is helping to increase sales for stores by allowing managers to understand the effects of weather and time of day. To put ourselves into the shoes of a customer, we would imagine that the predicted daily temp would be sufficient because people look at the weather once in the morning when planning out their day. If they see that the weather is predicted to rain, then they will not go shopping. If they are already at the store, and it starts raining, then they will likely continue shopping. For this reason, we recommend using the **predicted temperature_2** instead of the actuarial temp.

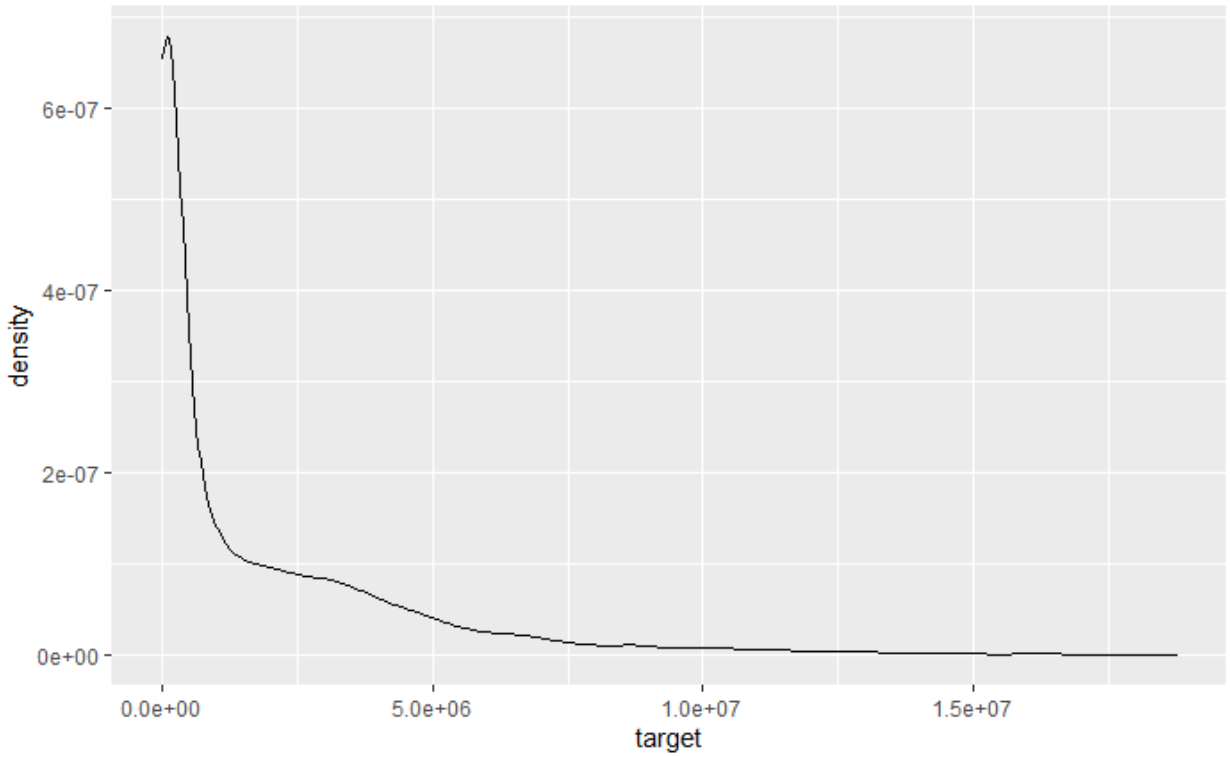
Task 4 – Consider transformations of the target variable (9 points)

A decision tree works by partitioning the feature space into rectangular regions and then predicting the mean of the target for each region. This asks a series of yes or no questions to each observation and based the answers it puts them into different buckets. The target variable **pedestrians** is right skewed and positive. Because a decision tree relies on the mean of each bucket, and means are sensitive to outliers, the decision tree will have better performance if we apply a skewness correction via a transformation prior to fitting the model.

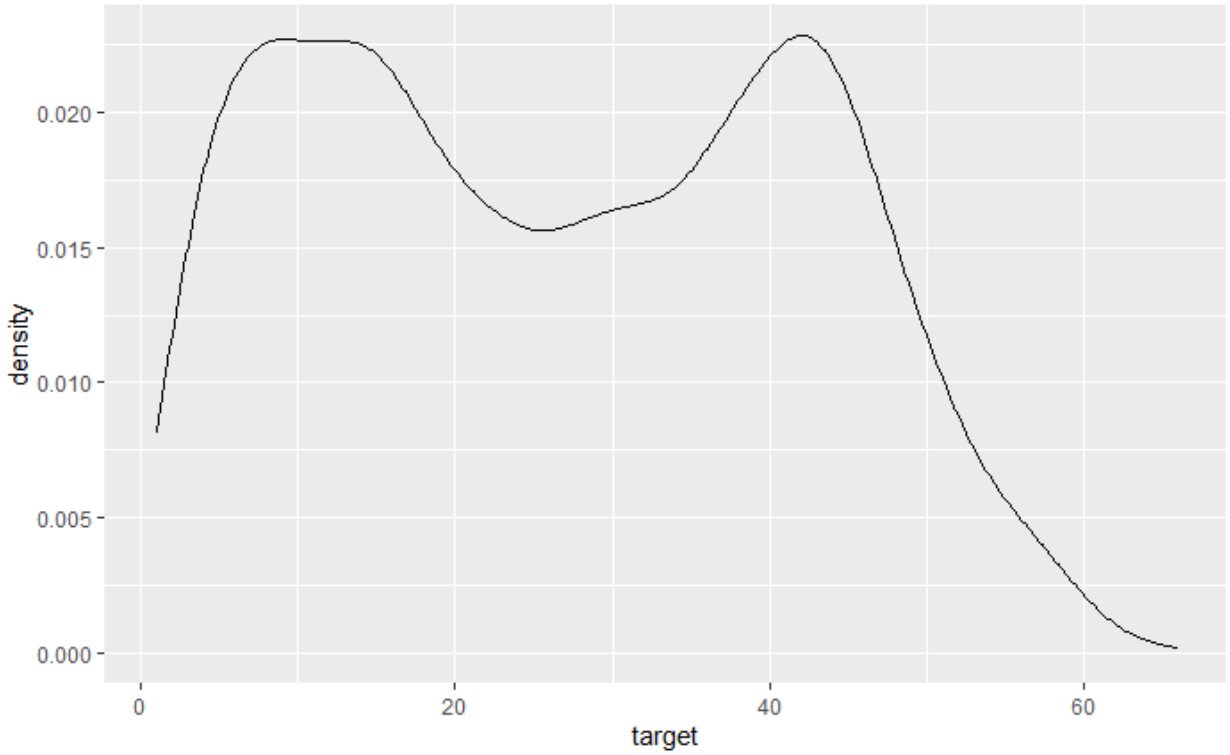




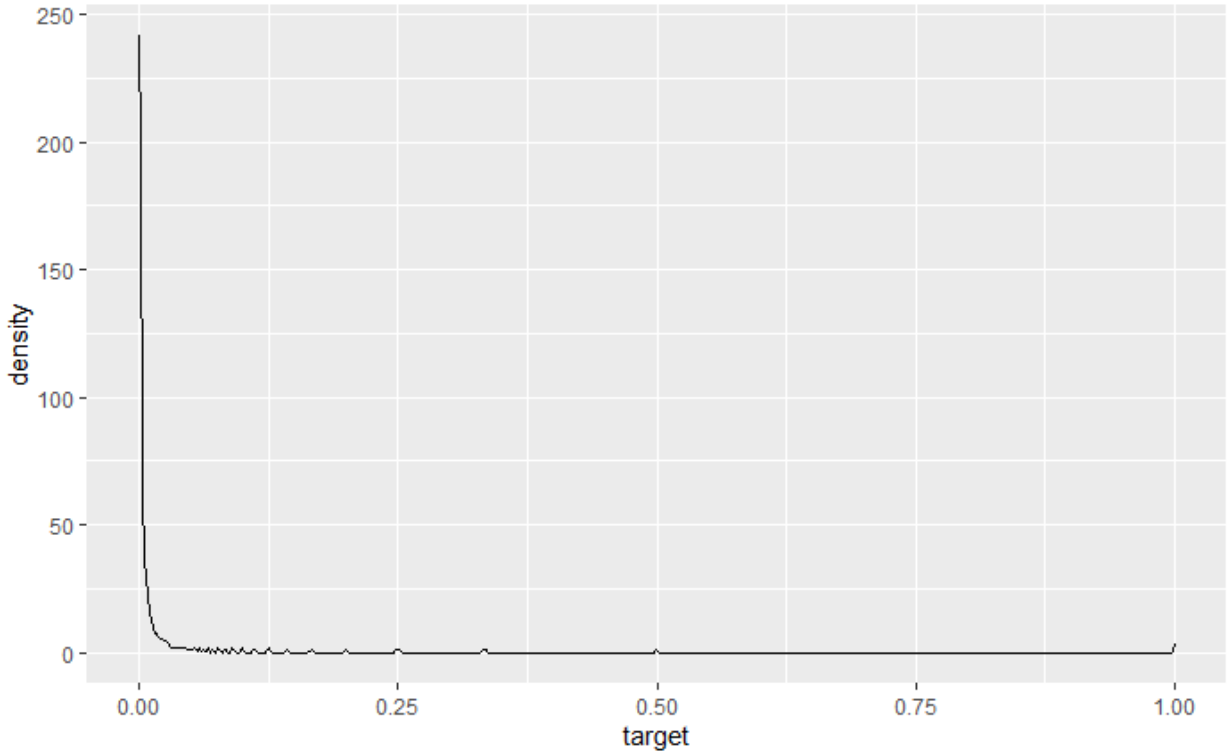
Taking the log only makes the skewness worse by causing a left skew instead of a right skew. We do not recommend this.



Taking the square causes the right skewness to increase. We do not recommend this.



The skewness is high when the variable is asymmetric and low when it is symmetric. The square root transform causes the distribution to be approximately symmetric between the values of 0 and 50 and a median that is equal to the mean of about 20. We recommend this as the transformation to use.



Taking the inverse is similar to taking the square in that the right skewness increases.

Task 5 – Build two trees (8 points)

In order to verify that our models will generalize into new data sets, we used a train and test set approach with a final holdout set. We first split the data into 30% and 70% which is used for training. Then we split the 30% into 10% which will be the final holdout and 20% which we will use to test the choices of our model parameters.

We do notice that the mean varies across these data sets. The validation set has the smallest sample size and also has a higher mean than the others. This is due to the random assignment of the records and so the results could change if the data were run using a different random seed.

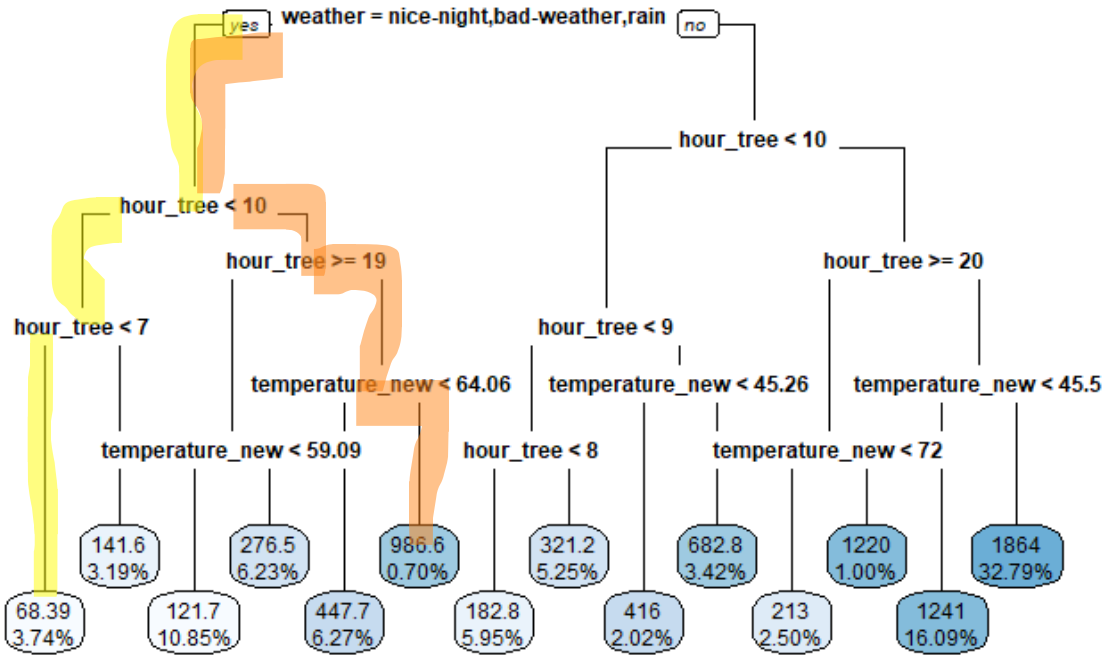
[1] "Mean value of pedestrians on data splits"

All records: 962.5301

Training: 960.577

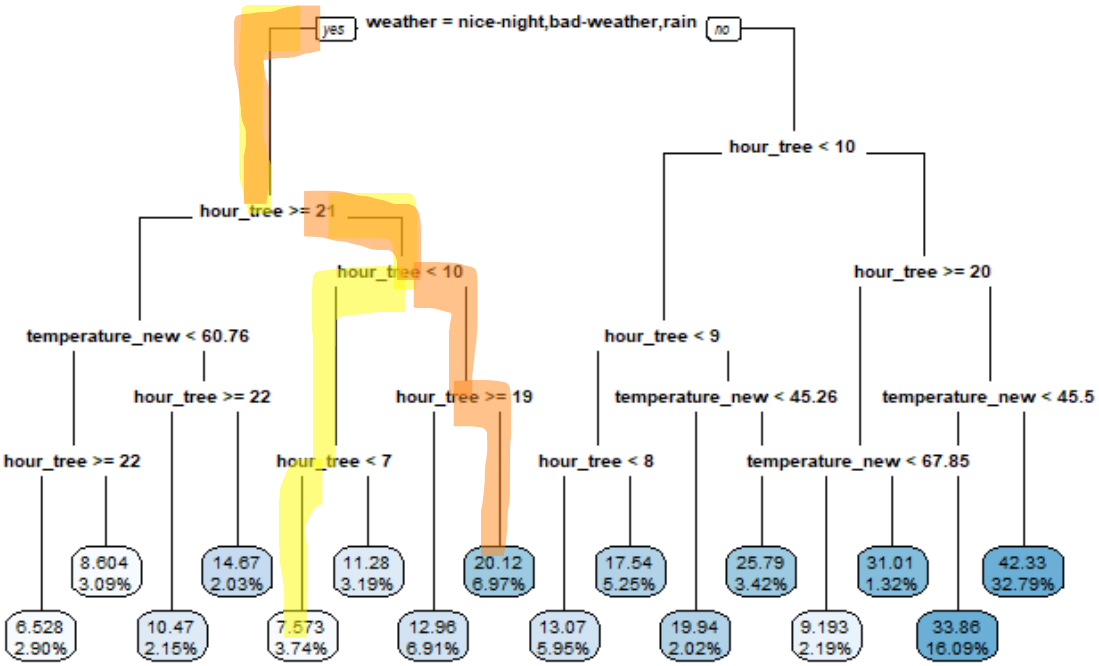
Testing: 962.741

Holdout Validation: 975.7984



The first tree makes these two predictions

- 68.4 pedestrians based on the hour being less than 7 and the weather being bad. This leaf represents 3.74% of the observations in the data.
- 986.6 pedestrians based on the hour being greater than 10 but less then 19 and the forecasted temperature being greater than 64. This represents 0.7% of the observations.



The second tree which has the square root transform makes these two predictions

- 7.573 because of the bad weather and hour. This represents 3.74% of the records. Notice that the square root is applied and so if we take the square, then we get 56.7 which is close to 68, the prediction of the first tree.
- 20.12, or when squared, 404.8 because of the hour and bad weather variables. This represents 7% of the records is lower than the prediction in the first tree.

The first tree uses the untransformed target and so the root mean squared error will be on a higher scale. For example, the first tree's predictions range from 60 to about 1,800 whereas the second only from 6.5 to 42. The RMSE takes the square root of the square of the residuals, and these will be greater for the first tree because the numbers are bigger. The assistant is wrong.

- RMSE of tree 1: 531
- RMSE of tree 2: 8.47

There are two ways that we could fix this problem by adjusting the error metric:

1. Take the square root of the pedestrians and the square root of the first tree
2. Take the square of the predictions from tree 2

We choose to use the first approach and note that the value is 8.579 which is higher (worse) than the second tree. The second approach leads to an value of 535.98 for tree 2 which is higher (worse) than for tree 1. These are conflicting results because they are not a one-to-one comparison.

Task 6 – Consider a random forest (3 points)

Random forests are ensembles of decision trees. In addition to more compute resources, random forests are more difficult to interpret because there are hundreds of trees used. Another consideration would be how to tune the hyper parameters for ntrees and mtry which can lead to overfitting because of the small size of the validation data set. We saw earlier that the mean of **pedestrians** was different in the validation set then in the training set, and so we spend a lot of time tuning a model and then applied it to a new data set, we risk overfitting and as a result having poor performance in real life. That is another consideration.

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

The Poisson distribution models a counting process. The **pedestrians** variable is the number of people who walk by during a given hour and so this can be understood as a Poisson process. This also matches the distribution because it is positive, discrete, and right skewed.

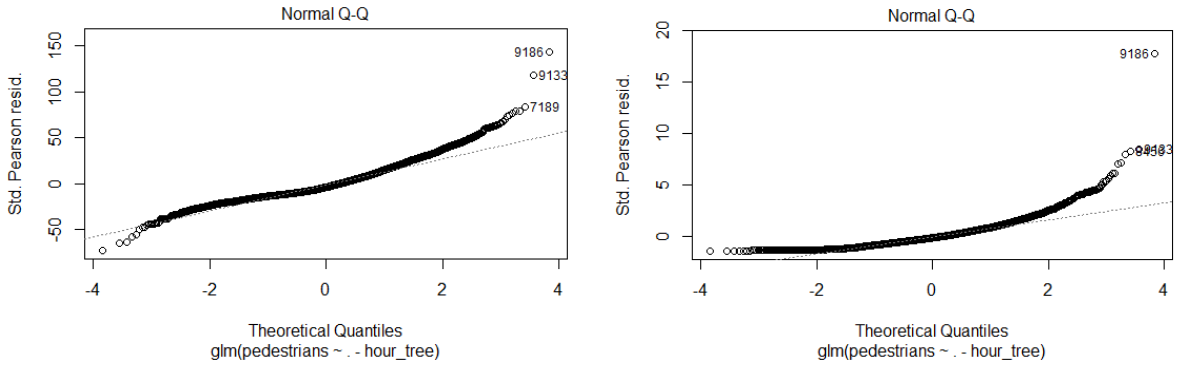
We could also use a Gamma or an Inverse Gaussian because these are positive and right skewed. They are continuous rather than discrete and so the predictions would come out as real numbers instead of positive integers. In this case, however, the client is only concerned about inference and does not need to use the model to make predictions.

```
Call:
glm(formula = pedestrians ~ . - hour_tree, family = poisson(link =
"log"),
    data = data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-99.237  -12.914   -3.903    7.618   71.953

Coefficients:
                   Estimate Std. Error  z value Pr(>|z|)
(Intercept)        7.152e+00  1.622e-03  4408.48   <2e-16 ***
weathernice-night -9.013e-01  1.820e-03  -495.34   <2e-16 ***
weatherbad-weather -9.323e-01  5.507e-03  -169.29   <2e-16 ***
weatherrain       -7.201e-01  3.409e-03  -211.21   <2e-16 ***
precipitation     -6.660e+00  5.158e-02  -129.12   <2e-16 ***
weekdayMonday     -8.588e-02  1.371e-03   -62.63   <2e-16 ***
weekdaySaturday    2.903e-01  1.248e-03   232.57   <2e-16 ***
weekdaySunday      7.514e-02  1.320e-03    56.92   <2e-16 ***
weekdayThursday   -1.291e-01  1.396e-03   -92.47   <2e-16 ***
weekdayTuesday    -1.629e-01  1.397e-03  -116.59   <2e-16 ***
weekdayWednesday  -1.805e-01  1.411e-03  -127.97   <2e-16 ***
hour_glm          -2.442e-01  1.861e-04 -1312.21   <2e-16 ***
temperature_new    1.255e-02  2.124e-05   590.88   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6814660  on 7962  degrees of freedom
Residual deviance: 1926449  on 7950  degrees of freedom
AIC: 1989523

Number of Fisher Scoring iterations: 5

[1] "Log link GLM 1 Train RMSE"
[1] 512.273
[1] "Log link GLM 1 Test RMSE"
[1] 519.4591
```

The log likelihood is a measure of how well the model fits the data. This is the log of the probability of generating these pedestrian counts conditional on the data. This is lower (519) than the GLM which has a Gamma response family (780), which implies that the Poisson is the best choice. It is worth noting that the second model is underfitting because the training error is worse (higher) than the testing error. Increasing the variance and decreasing the bias of this model by adding in more variables or creating new features would likely resolve this underfitting.

```
Call:
glm(formula = pedestrians ~ . - hour_tree, family = Gamma(link =
"log"),
    data = data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.9467  -0.5667  -0.1332   0.2998   4.4391

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         7.0417104  0.0351874 200.120  < 2e-16 ***
weathernice-night  -0.5540199  0.0228456 -24.251  < 2e-16 ***
weatherbad-weather -0.9327676  0.0681183 -13.693  < 2e-16 ***
weatherrain        -0.9677434  0.0443122 -21.839  < 2e-16 ***
precipitation      -3.2452542  0.4227574  -7.676 1.83e-14 ***
weekdayMonday      -0.0313416  0.0296691  -1.056 0.290830
weekdaySaturday     0.2514029  0.0297623   8.447  < 2e-16 ***
weekdaySunday       0.0360904  0.0296471   1.217 0.223513
weekdayThursday    -0.0440313  0.0296974  -1.483 0.138203
weekdayTuesday     -0.1116255  0.0296553  -3.764 0.000168 ***
weekdayWednesday   -0.0917453  0.0296718  -3.092 0.001995 **
hour_glm           -0.3688329  0.0037946 -97.199  < 2e-16 ***
temperature_new     0.0206395  0.0004598  44.893  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.497101)

    Null deviance: 12591.4  on 7962  degrees of freedom
Residual deviance:  4922.9  on 7950  degrees of freedom
AIC: 116388

Number of Fisher Scoring iterations: 7

[1] "Log link GLM 2 Train RMSE"
[1] 806.0683
[1] "Log link GLM 2 Test RMSE"
[1] 780.7006
```

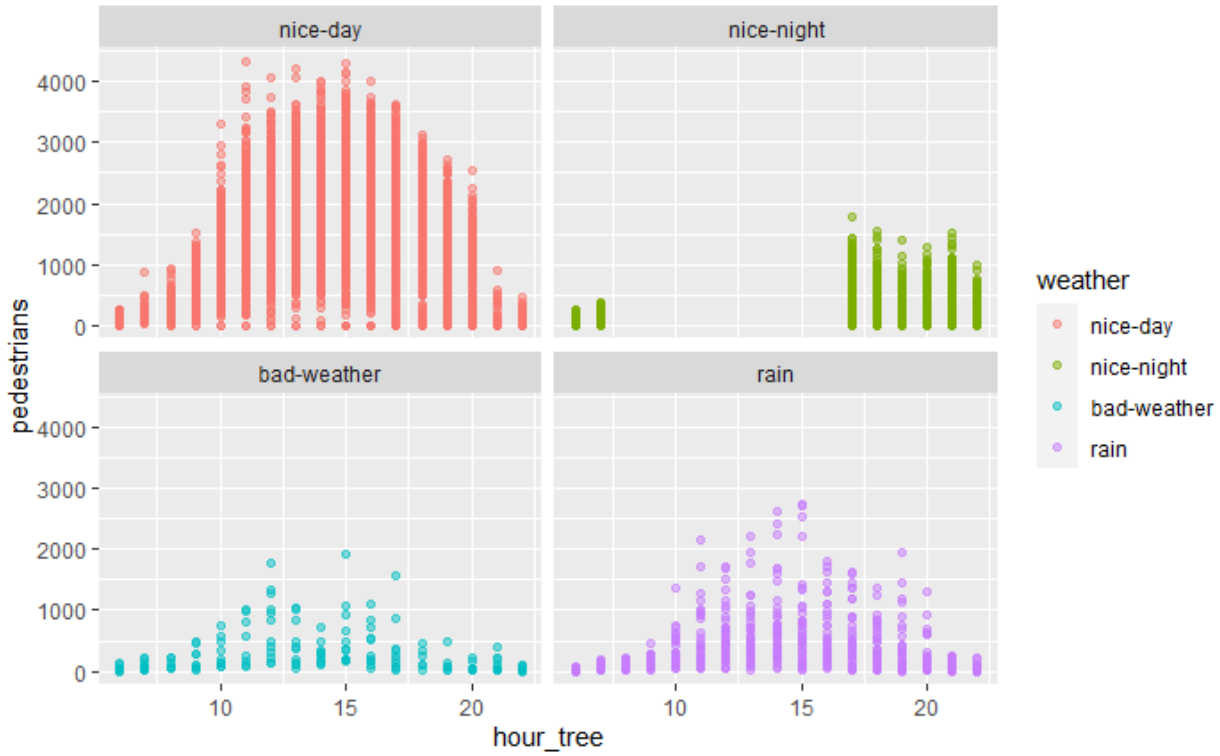
The business problem was to understand how time of day and weather impact sales. The second model has less certainty around the weekday coefficients, as is shown by the high p-values. This means that any results are subject to change if the data changes. The Poisson model has p-values that are all less than 0.001.



The GLM with a Poisson family (left) has deviance residuals which are closer to being normal than the GLM with the Gamma family (right). This shows that the Poisson model is a better fit to the data.

Task 8 – Consider an interaction (6 points)

An interaction is when the impact that a variable has on the target changes depending on the values of another variable that is in the model.



There is an interaction between hour and the weather. When deciding to go shopping, people will think about if they have enough time to get what they need before the store closes and if the weather is nice for walking. The upper left graph shows sunny and cloudy days. Here, there is a wide range of values because sometimes there is a lot of traffic and other times there is not. During bad weather or when it is raining, there are about the same number of pedestrians at any point in the day. This makes sense because during cloudy days it is not always easy to tell what time it is. The morning may be dark because of clouds and thunderstorms.

```
Call:
glm(formula = pedestrians ~ . + hour_glm * weather - hour_tree,
    family = poisson(link = "log"), data = data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-98.648  -12.415   -3.354    7.656   73.288

Coefficients:
                              Estimate Std. Error  z value Pr(>|z|)
(Intercept)                  7.123e+00  1.638e-03  4349.15   <2e-16 ***
weathernice-night           -1.339e-01  5.780e-03   -23.16   <2e-16 ***
weatherbad-weather          -7.764e-01  8.667e-03   -89.58   <2e-16 ***
weatherrain                 -5.120e-01  4.478e-03  -114.34   <2e-16 ***
precipitation               -6.563e+00  5.174e-02  -126.85   <2e-16 ***
weekdayMonday               -8.576e-02  1.371e-03   -62.55   <2e-16 ***
weekdaySaturday              2.911e-01  1.248e-03   233.16   <2e-16 ***
weekdaySunday                7.561e-02  1.320e-03    57.28   <2e-16 ***
weekdayThursday             -1.275e-01  1.396e-03   -91.34   <2e-16 ***
weekdayTuesday              -1.622e-01  1.397e-03  -116.04   <2e-16 ***
weekdayWednesday            -1.811e-01  1.411e-03  -128.39   <2e-16 ***
hour_glm                    -2.371e-01  1.917e-04 -1237.16   <2e-16 ***
temperature_new              1.274e-02  2.134e-05   597.11   <2e-16 ***
weathernice-night:hour_glm  -1.346e-01  9.826e-04  -136.96   <2e-16 ***
weatherbad-weather:hour_glm -5.689e-02  2.558e-03   -22.24   <2e-16 ***
weatherrain:hour_glm        -7.750e-02  1.158e-03   -66.92   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6814660  on 7962  degrees of freedom
Residual deviance: 1903536  on 7947  degrees of freedom
AIC: 1966615

Number of Fisher Scoring iterations: 5
```

This interaction is present in the data because the p-values on the **weathernice-night:hour_glm** and **weatherbad-weather:hour_glm** and **weatherrain:hour_glm** are all less than 0.001. If we had more time, then we would look at the RMSE of this model and see if it were lower than in the model without the interaction.

Interaction coefficient interpretation

- **During nice days –** there is no additional effect from the interaction with **weather** because “nice days” is the base level of the **weather** variable.
- **During nice nights –** The predicted count of pedestrians is multiplied by $e^{-0.1346} = 0.874$ for each unit increase in **hour_glm**, on top of the multiplies already introduced by the main effects. So the predicted count decreases by about 13% for each additional unit increase in **hour_glm**. The predicted count decreases at a faster rate per unit increase in **hour_glm** under “nice nights” than “nice days”. 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 hours.
- **When the weather is bad –** The predicted count of pedestrians is multiplied by $e^{-0.05689} = 0.944$ for each unit increase in **hour_glm.**
- **When it is raining –** The predicted count is multiplied by 0.925.

Task 9 – Select features (9 points)

Forward and backward selection are both methods of choosing which variables to include in the GLM based on the penalized log likelihood. When the log of the number of observations in the training data is greater than 2, BIC favors a simpler model than AIC does.

- Forward selection begins with no variables and only adds variables which lowers the BIC.
- Backward selection begins with all variables and only removes variables which lowers the BIC.
- Both directions continue iterating until the BIC stops getting lower. For Forward selection, it is likely that a simpler model will result because the BIC will stop improving while there are only a few variables. On the other hand, with backward direction, a complex model is more likely to result because the algorithm will stop as soon removing any variable will not decrease BIC.

We recall from the business problem that the goal is inference on the pedestrian activity and predictions do not need to be made. For this reason, we favor a simpler model which will be easy to interpret and so we recommend using the forward selection.

The algorithm begins with an intercept only model and then finds that adding **hour_glm** will decrease the BIC and so it adds this. Then it finds that **weather** will further decrease this and so adds **weather**. Then it continues until all variables are added.

```
Start:  AIC=6877716
pedestrians ~ 1

                Df Deviance     AIC
+ hour_glm       1  3297189 3360255
+ weather        3  4280125 4343208
+ temperature_new  1  6245615 6308681
+ precipitation  1  6540826 6603892
+ weekday        6  6575562 6638672
<none>              6814660 6877716

Step:  AIC=3360255
pedestrians ~ hour_glm

                Df Deviance     AIC
+ weather        3  2519505 2582597
+ temperature_new  1  2717045 2780119
+ precipitation  1  3037934 3101008
+ weekday        6  3060801 3123920
<none>              3297189 3360255

Step:  AIC=2582597
pedestrians ~ hour_glm + weather
```

```
Call:  glm(formula = pedestrians ~ hour_glm + weather + temperature_new +
    weekday + precipitation + hour_glm:weather, family = poisson(link = "log"),
    data = data_train)

Coefficients:
                  (Intercept)                      hour_glm            weathernice-night
                      7.12271                      -0.23712                      -0.13390
          weatherbad-weather                   weatherrain               temperature_new
                     -0.77640                      -0.51204                       0.01274
                weekdayMonday                weekdaySaturday                 weekdaySunday
                     -0.08576                       0.29109                       0.07561
              weekdayThursday                 weekdayTuesday               weekdayWednesday
                     -0.12749                      -0.16215                      -0.18114
               precipitation    hour_glm:weathernice-night  hour_glm:weatherbad-weather
                     -6.56279                      -0.13457                      -0.05689
        hour_glm:weatherrain
                     -0.07750

Degrees of Freedom: 7962 Total (i.e. Null);  7947 Residual
Null Deviance:      6815000
Residual Deviance: 1904000      AIC: 1967000
```

Task 10 – Recommend a model (8 points)

To decide on which GLM to use as the final GLM, we looked at the RMSE on the training and testing sets. This is different from using the log likelihood or AIC/BIC because these are only based on the testing set. The results below show that Task 7 shows that the Poisson family and log link are the best choices for response family and link function. Then in Task 8 we found that adding the interaction between **hour_glm** and **weather** improved the performance by reducing the error substantially. Finally, we confirmed that this is the best choice by running forward stepwise selection with BIC and returned the same model.

Task	Model	Train RMSE	Test RMSE
7	GLM with Poisson family and log link	512.273	519.459
8	GLM with Poisson family and log link with interaction	508.2897	516.618
9	Same from above	508.2897	516.618

The decision tree from Task 5 had higher RMSE indicating that the performance was worse than the GLM, at 531.597 instead of 516.618. For this reason, we recommend using the GLM. Although the objective is not to make predictions, having higher performance indicates that the results will be more representative to real life. A decision tree is easier to interpret, but if the tree is making bad estimates, then this just makes bad estimates easy to interpret and does not help the client to increase their retail sales.

Lastly, we retrained the GLM with the interaction over the entire data set. This has a larger sample size to estimate the coefficients and so the standard errors are smaller.

```
Call:
glm(formula = pedestrians ~ . + hour_glm * weather - hour_tree,
    family = poisson(link = "log"), data = data_all)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-98.567  -12.460   -3.509    7.672   71.928

Coefficients:
                                Estimate Std. Error  z value Pr(>|z|)
(Intercept)                    7.110e+00  1.371e-03  5186.25   <2e-16 ***
weathernice-night             -2.143e-01  4.881e-03   -43.91   <2e-16 ***
weatherbad-weather            -7.880e-01  7.326e-03  -107.56   <2e-16 ***
weatherrain                   -5.185e-01  3.678e-03  -140.98   <2e-16 ***
precipitation                 -6.099e+00  4.116e-02  -148.18   <2e-16 ***
weekdayMonday                 -6.911e-02  1.147e-03   -60.27   <2e-16 ***
weekdaySaturday                2.890e-01  1.051e-03   275.01   <2e-16 ***
weekdaySunday                  1.000e-01  1.104e-03    90.61   <2e-16 ***
weekdayThursday               -1.258e-01  1.171e-03  -107.44   <2e-16 ***
weekdayTuesday                -1.399e-01  1.165e-03  -120.12   <2e-16 ***
weekdayWednesday              -1.730e-01  1.186e-03  -145.91   <2e-16 ***
hour_glm                      -2.391e-01  1.607e-04 -1488.25   <2e-16 ***
temperature_new                1.289e-02  1.784e-05   722.56   <2e-16 ***
weathernice-night:hour_glm    -1.191e-01  8.256e-04  -144.24   <2e-16 ***
weatherbad-weather:hour_glm   -4.912e-02  2.171e-03   -22.62   <2e-16 ***
weatherrain:hour_glm          -8.193e-02  9.539e-04   -85.89   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 9768337  on 11372  degrees of freedom
Residual deviance: 2786661  on 11357  degrees of freedom
AIC: 2876730

Number of Fisher Scoring iterations: 5
```

Figure 4 Full data results in lower standard errors

Estimate	Std. Error
7.123e+00	1.638e-03
-1.339e-01	5.780e-03
-7.764e-01	8.667e-03
-5.120e-01	4.478e-03
-6.563e+00	5.174e-02
-8.576e-02	1.371e-03
2.911e-01	1.248e-03
7.561e-02	1.320e-03
-1.275e-01	1.396e-03
-1.622e-01	1.397e-03
-1.811e-01	1.411e-03
-2.371e-01	1.917e-04
1.274e-02	2.134e-05
-1.346e-01	9.826e-04
-5.689e-02	2.558e-03
-7.750e-02	1.158e-03

Figure 3 Training data results in higher standard error

Task 11 – Validate the model selection (4 points)

We already trained the model and chose the best parameters using the testing set. As a final step, we use the holdout set which represented 10% of the original data to test that our model works. This proves that in real life the results will be as effective as they are in this study.

The RMSE is 542.1039, which is higher than in the testing set. This is because we have already evaluated several models against the testing set and so it no longer was a blind validation. This small amount of overfitting is normal.

Task 12 – Executive summary (20 points)

To: National Retail Firm

From: Actuarial Consulting Firm

As your retail chain continues to expand an open more stores, we can help you to determine the most profitable locations. Your experts have informed us that weather is key for determining how much foot traffic these stores attract. For your existing stores, understanding how pedestrian activity varies by time of day and day of the week will allow you to optimize weekly ads, coupons, promotions, and staffing. Using predictive analytics, we present to you these findings which you can use to grow your business.

You can depend on the accuracy of these results. We conducted a data analysis using data from the City of New York from 2017-2019. With 11,373 observations of the number of pedestrians in each hour, we looked at the hour of the day, precipitation, forecasted temperature, actual temperature, weather conditions, and weekday. We inspected this data for errors and found none.

Data Exploration

Humans are habitual creatures who shop during certain days of the week; however, the day of the week alone was not sufficient to guarantee if there would be a lot of traffic or not. The time of day is closely related to the number of pedestrians who are out. You may like to know that during 10 am – 6 pm are the highest traffic times of the day.

Humans like things to be simple. Weather is a complex phenomenon, and you can benefit by simplifying every day down into four categories: a "nice day", a "nice night", "bad weather", or "rain". Most people go out in fair weather and stay indoors during poor weather.

The time variable is critical to helping you to maximize revenue for your existing stores. When do people shop, that is the key question. We considered three ways of thinking about time:

- **Clock time**: this is the exact time that we see on the clock, from 8 am in the morning to 10 pm at night.
- **Time blocks**: these hours of time such as "11 am brunch", "12 pm lunch", "1 pm coffee hour", and so forth.
- **Relative time**: the number of hours before or after 2 pm, which is the peak time.

Predictive Modeling

Depending on the type of model, we recommend using either the first or the last measure. The first has the advantage of being easy to understand and would work well for a decision tree model. This is a series of yes/no questions which you would ask about the weather situation, day of the week, hour, and so forth. The last measure is useful for linear models where you need an approximate way of measuring time.

Shopping and spending are related to our emotions. How we feel based on the music in the store, and the temperature outside does impact our spending habits. Cold weather keeps people indoors while warm weather encourages people to go outside. The forecasted temp is a better indicator of traffic than the actual temp and so we recommend looking at weather forecasts when choosing your new store locations. A word of caution is to also consider the other factors at play. Using temperature alone is not sufficient for predicting the number of pedestrians who are out and about.

As those who work in retail are familiar, when it rains it pours when it comes to customer traffic. Days when it is slow seem to drag on for weeks on end whereas days when it is busy fly by quickly. We considered this feeling in our models by adjusting the number of people who were active. If just used the count of people by themselves, we would run into problems because when there are 10,000,000 people outside then an additional 1,000 people is not surprising; however, if there are only 50 people outside, it is extremely unlikely that an additional 1,000 would come about unexpectedly. Mathematically, this is known as a skewed distribution. To adjust for this, we tested out different formulas for compensating for this. We found that taking the square of the number of people caused this measure to be more stable statistically. We did not use this in our result but are including it here for completeness.

Every situation is unique and requires a different type of solution. There are many types of models to choose from and each have their advantages and disadvantages. For your problem, because you are looking for simple and easy-to-understand results, we tested out generalized linear models (GLMs) and decision trees.

The decision tree is a series of yes or no questions which can be thought of as a series of rules that get applied to every row in the data. On any given hour, the tree asks what the weather is, what the temperature is, if it is raining or not, and so forth, and based on this it comes up with an estimated number of pedestrians who are about during these types of days and times. We fit two different trees and chose the one which most closely matched the data.

The GLM relates the number of pedestrians to many variables at once so that we can take into consideration the exact values of the temperature, time of day, weather, and so forth. Then it uses a formula to calculate several pedestrians. We fit three different GLMs and chose the one which did the best job of matching the data.

We did not consider more advanced machine learning methods because of two concerns that you would likely have are:

1) The amount of computing resources and IT infrastructure required.
2) The possibility of being wrong by miscalculating a result. Complex models require more attention from the actuary or data scientist who is building and maintaining them.

Results

We used a common statistical measure of performance known as the root mean squared error (RMSE). When this is high, the model is a poor fit and when it is low it is a good fit. We can be assured that these results will generalize in real life because we designed an experiment using validation testing. This keeps part of the data in a blind holdout set that our actuaries did not look at until the very end of our report. Once we had chosen our model, we ran this data through our model and verified that the predictions matched the actual pedestrian activity.

The GLM had the best performance out of the three GLMs and two decision trees tested. We also considered how easily you would be able to use the result, and as you will see below, this model below gives you invaluable insights into pedestrian activity. The formula will replicate the results of our model within a spreadsheet. You can translate this just by starting at the top row and moving downwards.

Interpretation
Start with 1,240 pedestrians on a nice Friday
If it is night, multiply by 0.87
If the weather is bad, multiply by 0.46
If it is raining, multiply by 0.6
Multiply by 0.001^n where n is the number of inches of precipitation
If Monday, multiply by 0.92
If Saturday, multiply by 1.34
If Sunday, multiply by 1.08
If Thursday, multiply by 0.88
If Tuesday, multiply by 0.85
If Wednesday, multiply by 0.83
Multiply by 1.01^n where n is the temperature
If it is a nice day, multiply by 0.79 for each hour past 2 pm
If it is a nice night, multiply by 0.69 for each hour past 2 pm
If it is bad weather, multiply by 0.75 for each hour past 2 pm
If it is raining, multiply by 0.73 for each hour past 2 pm

This provides us with useful insights such as

- The most common day to be outside is on Friday. This is the end of the work week for many people and so they want to get out of the office and spend time outdoors.
- Streetlights and illumination should be considered in the new locations. Pedestrians do not walk on dimly lit sidewalks and so to continue attracting customers during nighttime hours, we recommend installing light fixtures around the stores.
- Weather needs to be considered. When designing buildings in cities that experience bad weather, care should be taken to allow for covered entrance ways so that pedestrians will come into the stores for shelter. We advise against marketing to walk-in customers in areas where there is continual rainfall throughout the year.
- The day of the week is going to be the same regardless of the location. We did not see any difference in the weekday during different temperature levels. We recommend launching shopping marketing campaigns Starting on Friday and into Sunday because the multipliers are highest here, from 1.00 on Fri, 1.34 on Sat, and 1.08 on Sun.
- Time depends on the weather. If it is a nice day, pedestrians are in less of a rush and are more willing to stay around outside.
- Caution needs to be taken when interpreting the third-to-last entry, which refers to the nice night and time relationship. Clearly, when it is night then it is past 2 pm already.

Conclusion and Next Steps

This result shows that pedestrian activity is not random and is based on known and quantifiable conditions which are publicly available. The data was only from New York and so in other states with different weather patterns and populations these results would necessarily be different. A thorough updating on new data should be conducted. There were only a few variables used and so there is

information that is not being included such as whether there is a COVID pandemic that is enforcing social distancing, the population of the area, or whether it is a city or suburban location.