Statistical Analysis of New York City Yellow Taxi Ridership

Geng Yuxiang

September 9, 2019

Abstract

As peer-to-peer ridesharing platform like Uber and Lyft start taking over the taxi market in New York[1], it is important for taxi companies to have knowledge of the factors that could affect profitability in order to make better decisions in taxi-task assignment.

1 Introduction

In this study, I explored 5 predictor variables that could potentially affect earnings. In particular, I studied how these predictors could affect total amount earned in any given hour. The 5 predictors include

PULocationID: Taxi zone where the trip began

Day of the Week: Monday to Sunday

Hour of the Day: e.g. 7am, 9pm

Weather Condition: State of the weather (e.g. Heavy snow, light rain)

Temperature: e.g. 17°C

And the response variable is

Total Amount (Hourly): Total Amount earned in hourly basis

The goal of this study is to construct models that could make reasonable predictions in a scenario such as what is the total earning if trips are taken from 5 pm - 6 pm on a rainy Thursday and the temperature is 17°C.

2 Data Selection and Tools used

There are 2 data sets used for this study: taxi data and weather data.

The taxi data set consists of yellow taxi data selected from the entire year of 2017. Since it would be time-consuming and memory consuming to process the entire year of data, I randomly selected 100k data points from each month and combined into a single data set. This also helps reduces bias from each month.

The weather data set was scraped from timeanddate.com[1] with BeautifulSoup[2] and Selenium Webdriver[3], which includes hourly weather information (temperature, conditions, wind, etc.) for the whole year of 2017.

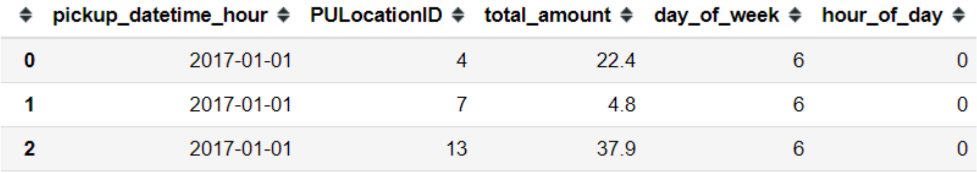
[1] https://www.timeanddate.com/weather/usa/new-york/historic?month=1&year=2017

[2] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[3] https://www.seleniumhq.org/

3 Preprocessing and Cleaning

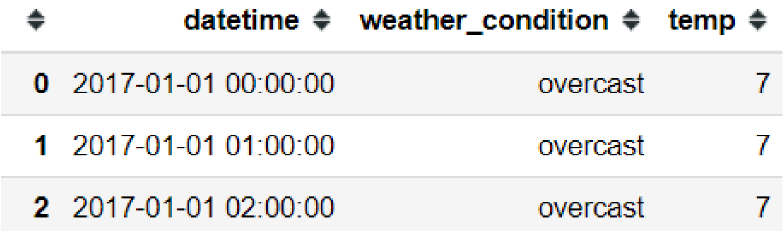
For the taxi data set, I first removed 2 instances of 2008 and 2009 data points which were most likely anomalies as they are not 2017 data. Also, instances with negative total_amount were removed. Then, I isolated tpep_pickup_datetime, PULocationID (factor) and total_amount (double) as they are my variables of interest. Next, since my goal is to compare against hourly total amount, trips that took place in the same hours are considered equivalent. Therefore, tpep_pickup_datetime were rounded down to the nearest hour (e.g. from 2017-2-14, 13:21:00 to 2017-2-14, 13:00:00). Next, since both time and location (taxi zone, in this case) are considered in this study (and they are independent), I will need hourly total amount with respect to every combination of pick-up time and taxi zone. This is achieved by grouping tpep_pickup_datetime and PULocationID simultaneously and taking sum of total amount. Next, I extracted hour_of_the_day (factor) and day_of__the_week (factor) from each tpep_pickup_datetime.



⬍	pickup_datetime_hour ⬍	PULocationID ⬍	total_amount ⬍	day_of_week ⬍	hour_of_day ⬍
0	2017-01-01	4	22.4	6	0
1	2017-01-01	7	4.8	6	0
2	2017-01-01	13	37.9	6	0

Figure 1: First 3 rows of taxi data set after preprocessing

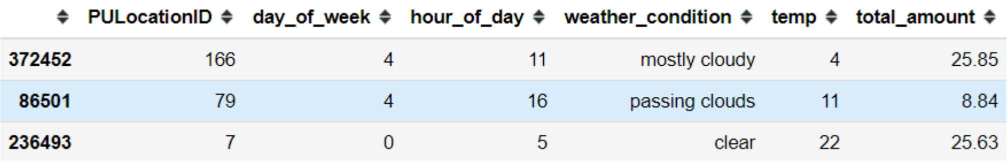
Next, I preprocessed the weather data set. Each instance consist of temperature, weather, wind and some other information. Each instance was recorded hourly at the 51st minute (e.g. 12:51 am). This would not match tpep_pickup_datetime from taxi data set. Therefore, recorded_datetime were rounded to the nearest hour. I chose to round up (e.g. from 2017-2-14, 13:51:00 to 2017-2-14, 14:00:00) instead of rounding down, which could otherwise lose consistency due to longer time interval. Additional, there were missing records at certain hour of the day. This is handled by backward filling to propagate the next values backward. Finally, I separated recorded_datetime, which will be used for joining with taxi data, and temperature (integer) and weather_condition (factor), which are 2 of our predictor variables.



⬍	datetime ⬍	weather_condition ⬍	temp ⬍
0	2017-01-01 00:00:00	overcast	7
1	2017-01-01 01:00:00	overcast	7
2	2017-01-01 02:00:00	overcast	7

Figure 2: First 3 rows of weather data set after preprocessing

At last, the two data sets were joined by datetime and the joined data set will be used later in statistical analysis. The response variable is total_amount (Continuous) and the 5 predictor variables are PULocationID (Factor), day_of__the_week (Factor), hour_of_the_day (Factor), weather_condition (Factor) and temp (Discrete). Here are 3 randomly selected rows of the final data set.



⬍	PULocationID ⬍	day_of_week ⬍	hour_of_day ⬍	weather_condition ⬍	temp ⬍	total_amount ⬍
372452	166	4	11	mostly cloudy	4	25.85
86501	79	4	16	passing clouds	11	8.84
236493	7	0	5	clear	22	25.63

Figure 3: 3 randomly selected rows of the final data set

4 Descriptive Statistics

In this section, I explored distribution of variables and correlation between them.

4.1 Distribution

First, I looked at distribution of the response variable total_amount. A sample of 10k is randomly selected as visualizing larger sample is time-consuming and patterns are not distinctive. The first plot shows the distribution as histogram. We can see that the majority of total amount earned hourly are between 10 and 50 and density decreases as total amount increases. The sample mean is calculated to be 49.56042 and standard deviation 55.54739. The distribution appears to be right-skewed and therefore Weibull, Log Normal and Gamma distributions would be more fitting. Therefore, I fit the sample with these three distributions using method of maximum likelihood estimation(MLE). From the fitting curves in the first plot, the differences between the three are insignificant. However, the rest of the three plots shows that Gamma distribution is the most appropriate. The CDF plot shows that the fitted Gamma CDF are mostly align with CDF from sample whereas the other two are slightly off. The Q-Q plot shows that sample points are mostly align with Gamma Q-Q plot with slight trait of heavy tails on upper portion. But this is not as significant as the other two distributions. As for P-P plot, sample probability is more align with probability from gamma distribution as compared to the others. Overall, gamma distribution best fits the data. Also, the estimated parameters from MLE are shape = 1.29820270 and rate = 0.02620458.

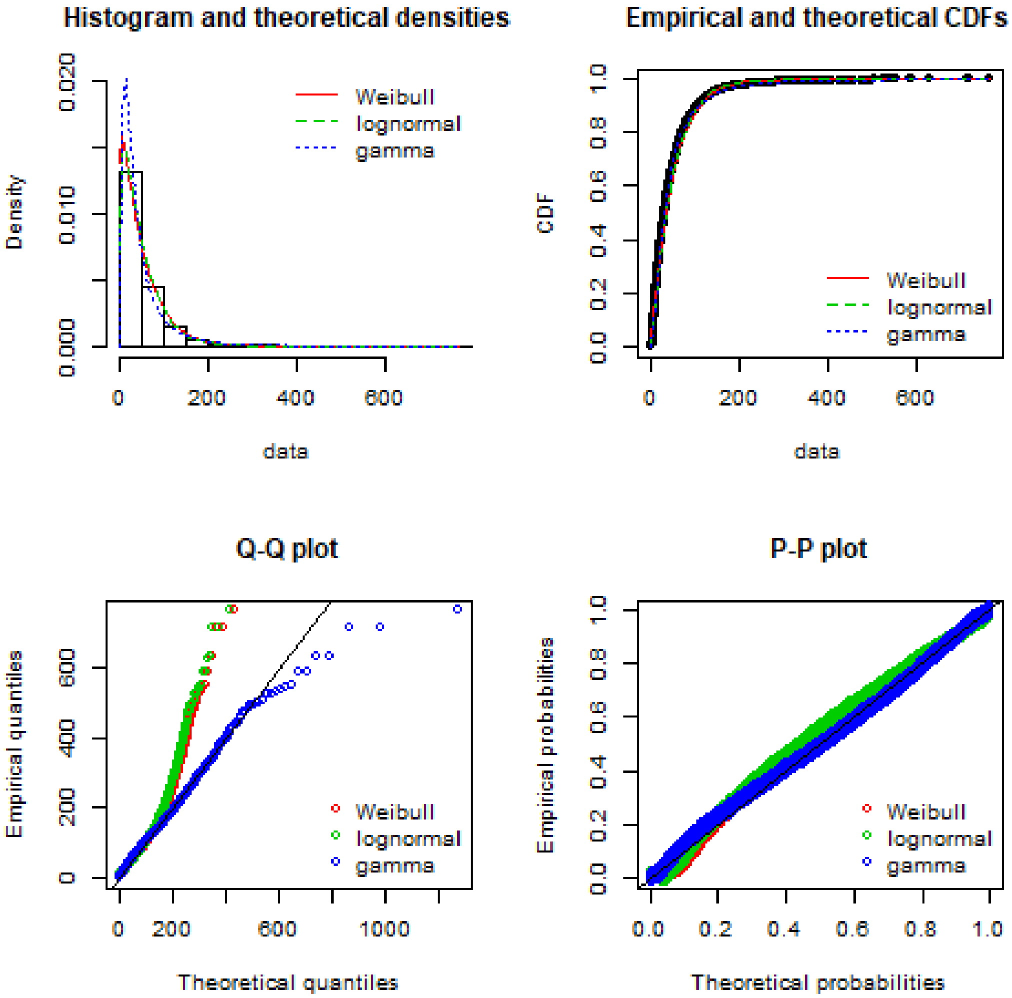


Figure 4: Total Amount Distribution

Next, I looked at distribution of temperature. A sample of 10k is again used for fitting distribution. The histogram shows that majority of temperatures (hourly) are between 5°C and 25°C. Density/Probability of occurrences are lower on both tails. The sample mean is calculated to be 13.054 and standard deviation 9.725002. The distribution generally appears to be normal. I fit a normal curve generated from MLE and it generally fit the distribution well. The sample CDF is align with CDF from generated normal. The Q-Q plot shows sign of light tails but generally appropriate. The sample probability is mostly align with normal probability. Note that since temperature are whole integers, theoretical quantiles may take up a small range for each sample quantiles. Overall, normal distribution fits the data and the estimated parameters from MLE are mean = 13.054000 and standard deviation = 9.720138.

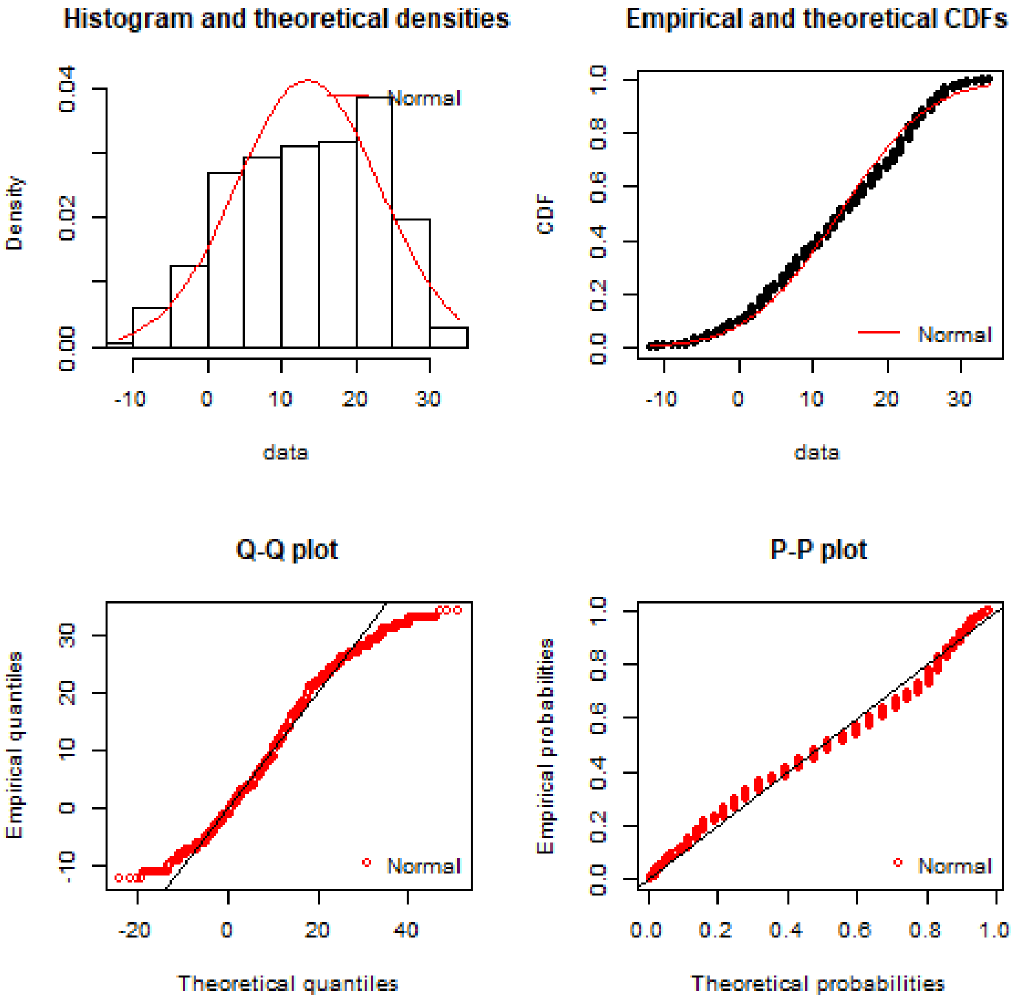


Figure 5: Temperature Distribution

4.2 Correlation between Response and Predictor Variables

Next, I looked at correlation between total_amount and each predictor variables individually. In this case, 10k sample points are randomly selected as large sample makes patterns indistinctive.

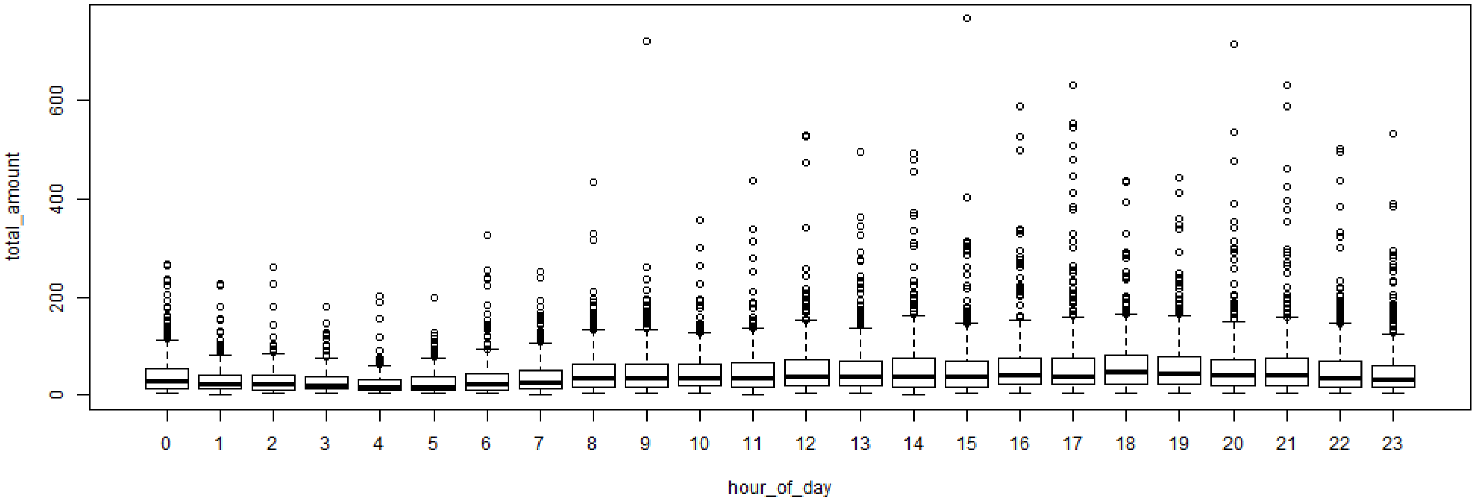


Figure 6: total amount vs hour of day

This plot shows total amount in every hour of the day. There are distinctive variation in different hours. It appears that total amount are higher from 7 to 21 and lower in other hours. This could be because of higher demand for taxi during daytime, especially during peak hours.

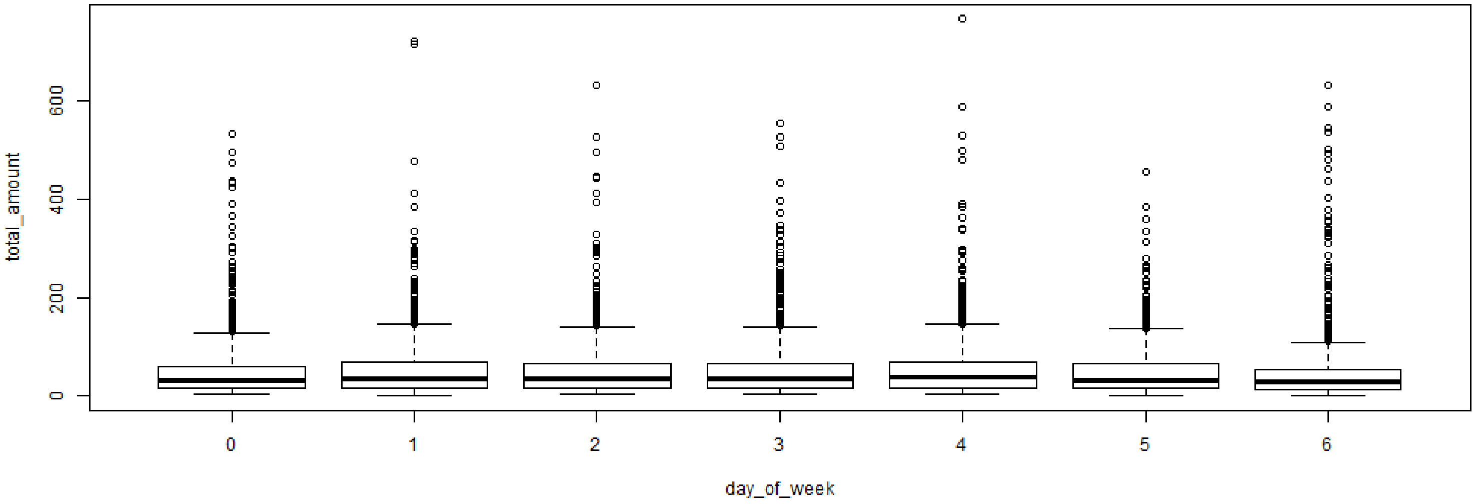


Figure 7: total amount vs day of week

This plot shows total amount in every day of the week. Note that weeks start from 0 (i.e. 0 is Monday). It appears that there is a little variation but not apparent to the eyes. Therefore, it is necessary to check for its statistical significance, which is done later on.

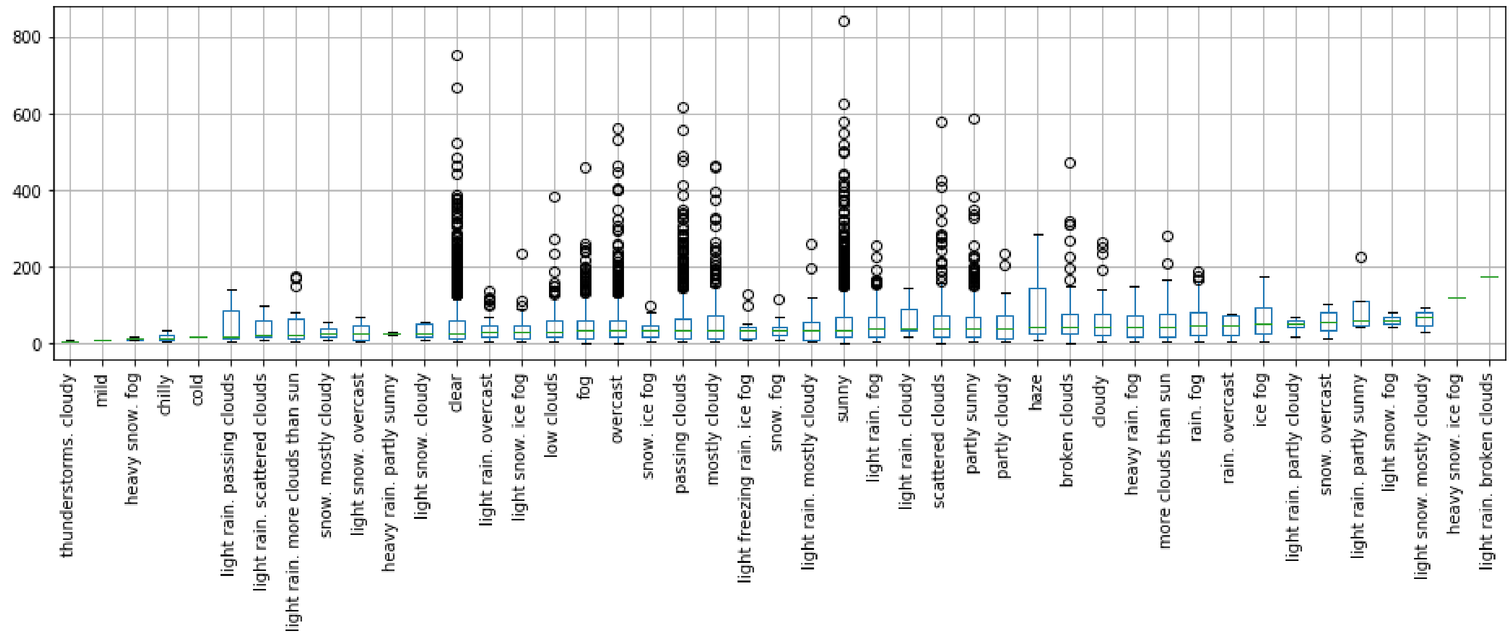


Figure 8: total amount vs weather condition

This plot shows total amount in different weather condition. There appears to be large variation between each condition. I can see that total amount is greater when there is light rain, light snow, etc. and smaller when there is broken cloud and chilly, etc.

4.3 Test for Significance of Parameters

Next, I perform Analysis of Variance (ANOVA) on each predictor variable to check if there is difference between factor levels. It turns out that PULocationID, day_of_the_week, hour_of_the_day are strongly significant with p value of <2.2e-16, 2.221e-05, <2.2e-16 respectively. Also, temperature is significant with p value of 0.001638. However, weather condition turns out to be less significant with p value of 0.1168.

4.4 Check for Correlation between Predictor Variables

It is natural to think that temperature could be related to weather condition. For example, it is usually colder when it is snowing. Also, this study[2] on The Scientific World Journal concluded that there is negative correlation between precipitation and temperature. Also, box plot of temperature against weather condition which I constructed shows that temperature is lower at snowy days and higher when it is rainy. Hence, I may need to consider interaction in model fitting later on.

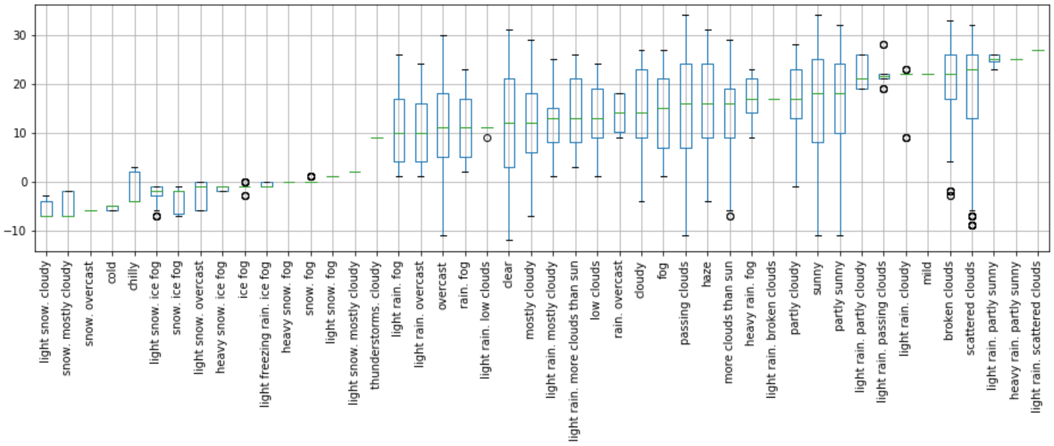


Figure 9: Temperature vs Weather Condition

5 Statistical Modelling

In the section, I will walk through steps taken to select the best models that fit the data. 300k data points are randomly selected for model fitting as larger data set usually makes our estimates more precise.

5.1 Model Fitting and Transformations

I fitted a linear model (Linear Model 1) to the data with all 5 predictor variables against total_amount. This is equivalent to a normal regression with identity link function. To check the fitting of model and linear model assumptions are fulfilled, I need to look at residual plot.

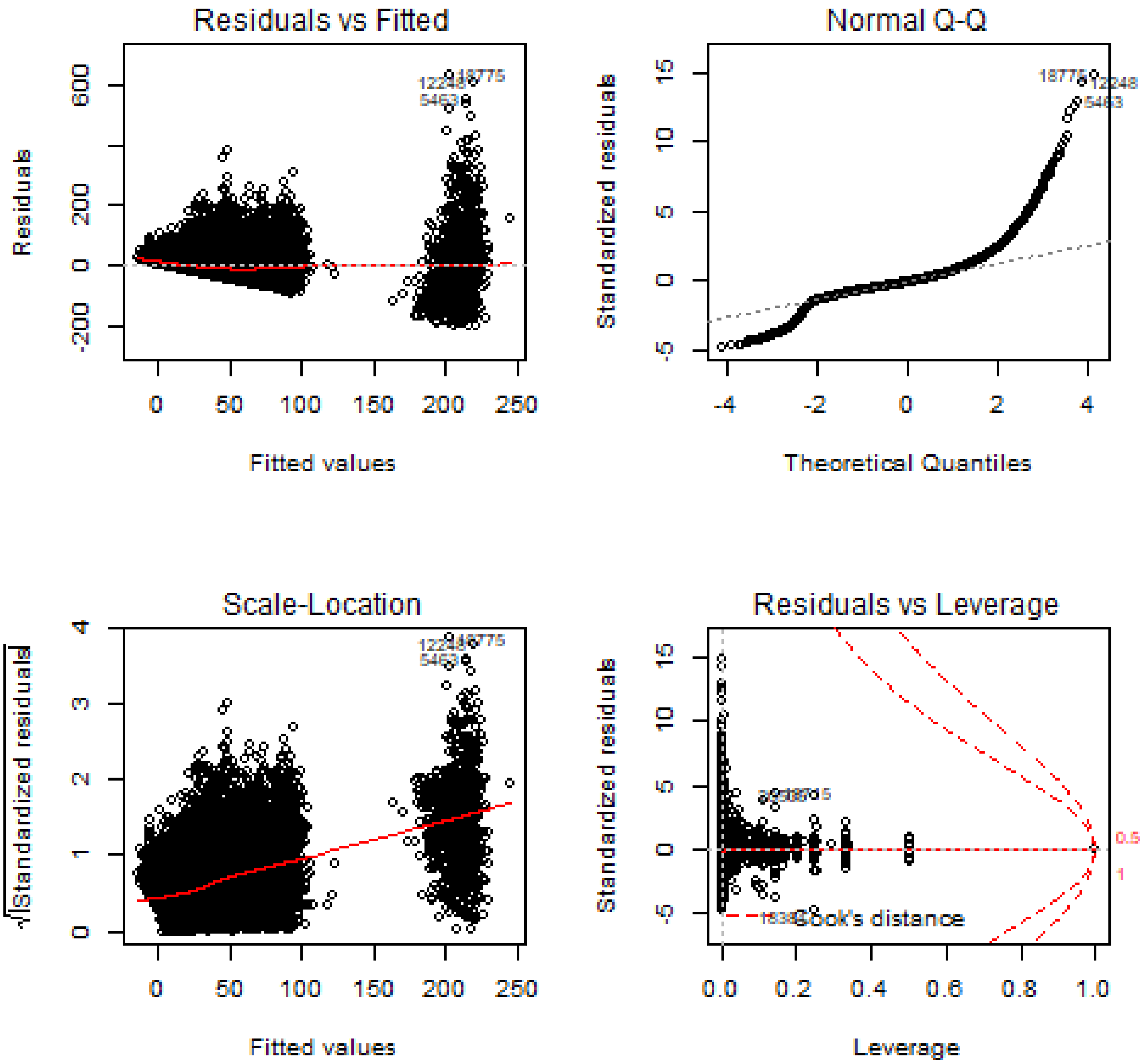


Figure 10: Linear Model 1

The Residuals vs Fitted values plot shows that variance of residuals increases as fitted values increase. This indicated heteroscedasticity in the model, which is reaffirmed by Standardized Residual vs Fitted Value plot, where the fitted line has a positive gradient and spread of standardized residual increases. These traits indicates the need to transform our model variables. In particular, we need to scale down response variable to reduce heteroscedasticity. The Normal Q-Q plot have heavy tails, which indicates error is not quite normally distributed. This violates the assumption of linear model and should also be dealt with transformation. Next, I transformed the model by taking log of total_amount and constructed Linear Model 2.

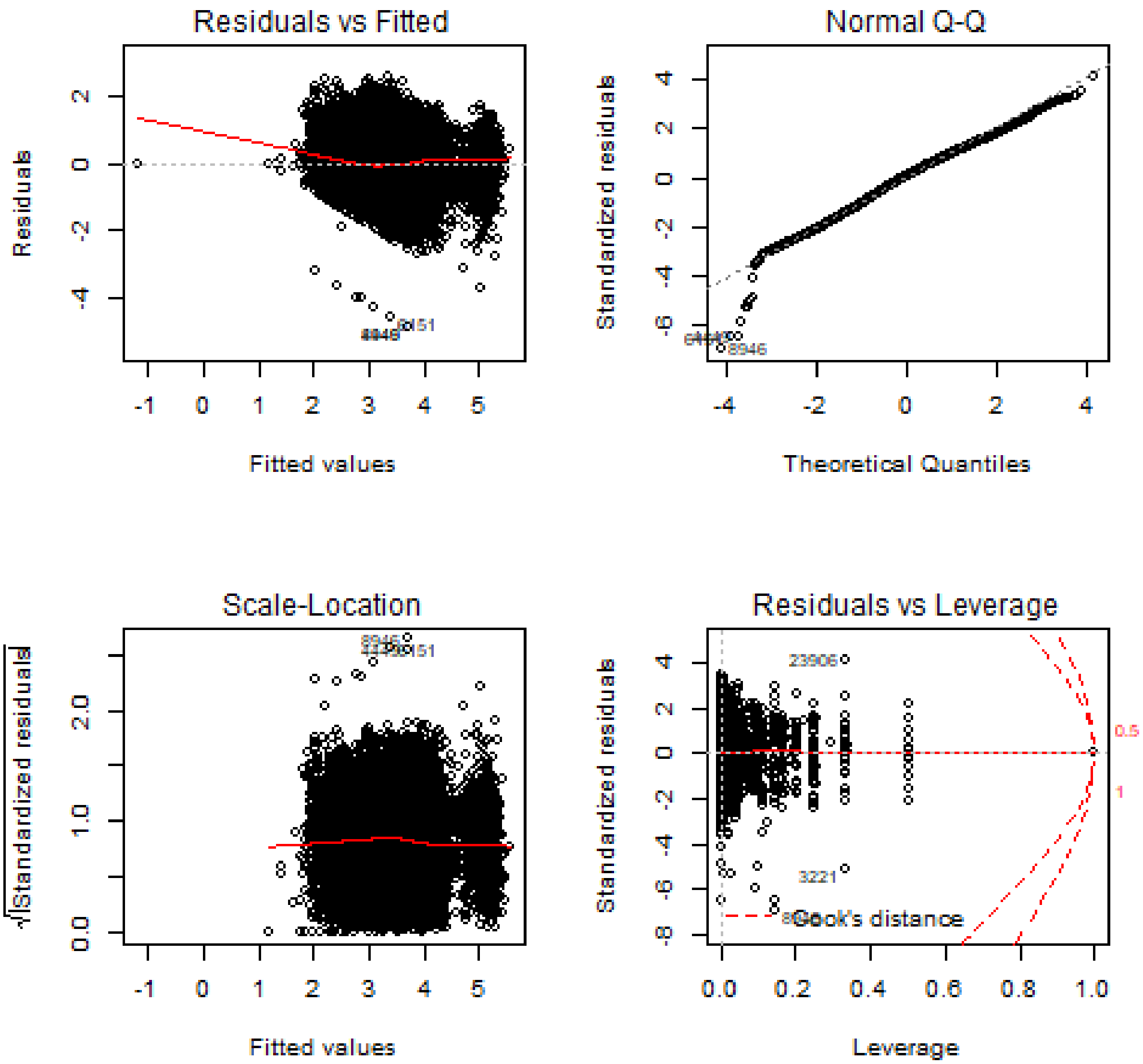
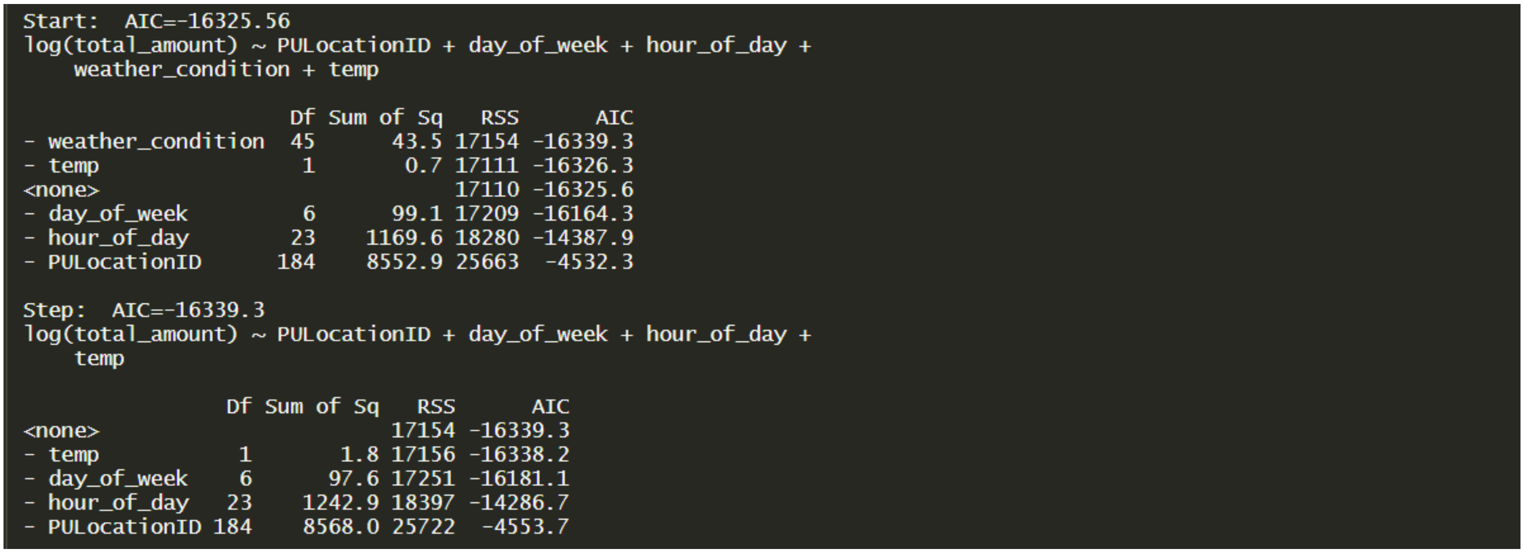


Figure 11: Linear Model 2

Overall, taking log transformation appears to be effective. The Residuals vs Fitted values plot shows that residuals have mean of 0 and evenly distributed along 0 mean with no fixed pattern, indicating consistent variance. This is reaffirmed by Standardized Residual vs Fitted Value plot which shows that standardized residuals are evenly distributed along consistent mean. The Normal Q-Q plot shows almost perfect fit, indicating that error with log transformation is normal distributed. Also, the Residual vs Leverage plot shows few points with high leverage or residual. There are a few points has leverage or high residual but their counterpart is rather small, making cook's distance quite reasonable).

5.2 Model Selection

Next, I applied stepwise selection to linear model 2 eliminate parameters that are less significant in the presence of other parameters.



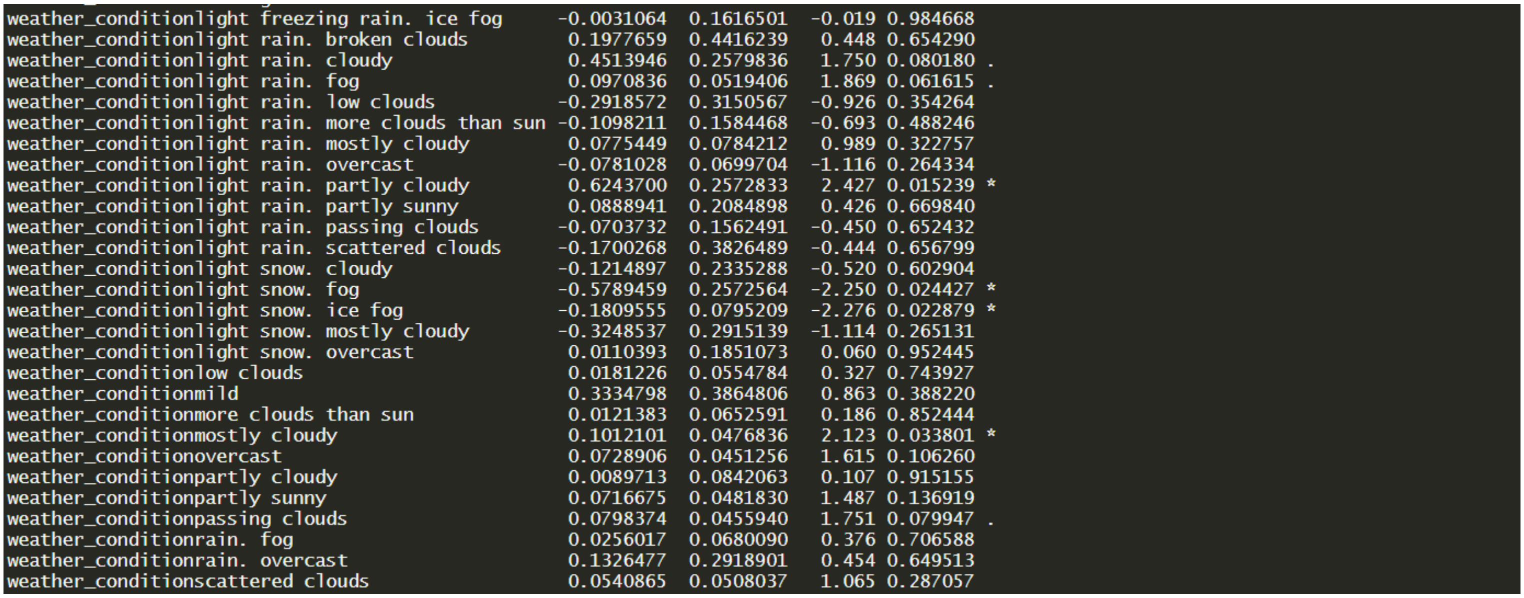
```
Start:  AIC=-16325.56
log(total_amount) ~ PULocationID + day_of_week + hour_of_day +
    weather_condition + temp

                    Df Sum of Sq   RSS      AIC
- weather_condition  45      43.5 17154 -16339.3
- temp               1       0.7 17111 -16326.3
<none>                            17110 -16325.6
- day_of_week        6      99.1 17209 -16164.3
- hour_of_day       23    1169.6 18280 -14387.9
- PULocationID     184    8552.9 25663  -4532.3

Step:  AIC=-16339.3
log(total_amount) ~ PULocationID + day_of_week + hour_of_day +
    temp

               Df Sum of Sq   RSS      AIC
<none>                      17154 -16339.3
- temp          1       1.8 17156 -16338.2
- day_of_week   6      97.6 17251 -16181.1
- hour_of_day  23    1242.9 18397 -14286.7
- PULocationID 184    8568.0 25722  -4553.7
```

Figure 12: Stepwise 1

The stepwise selection dropped weather_condition. From ANOVA result from last section, we know that weather_condition is not as significant as other variables. This is most likely also true in the presence of other variables. The stepwise selection improves the fit by decreasing AIC from 68812.75 to 68799.01 by dropping weather_condition.

5.3 Refine Model: Is Weather Condition Useless At All?

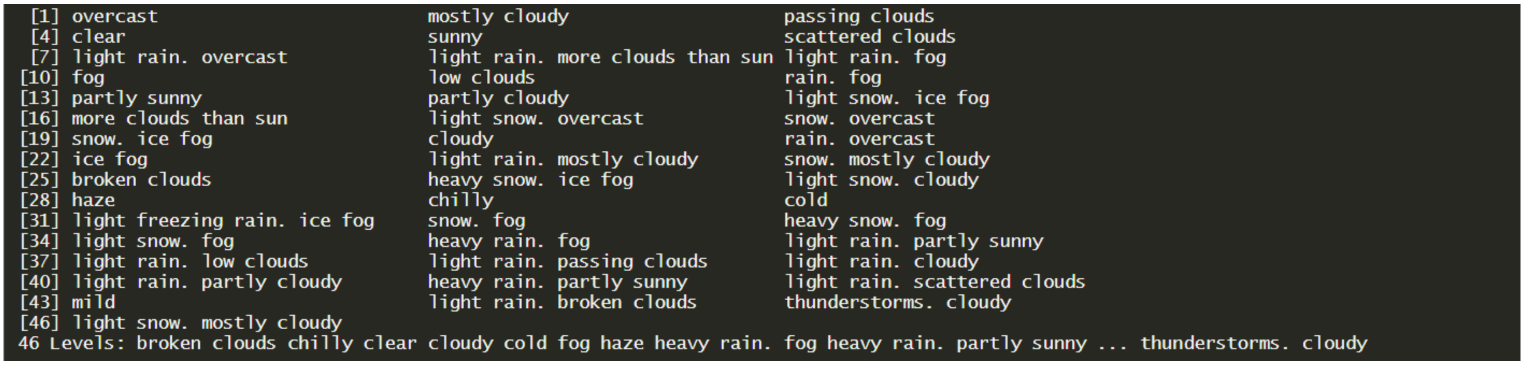
From model summary, most weather_conditions are considered insignificant in the presence of other variables, which suggests most weather conditions have little impact on changing total amount earned in an hour.



```
weather_conditionlight freezing rain. ice fog     -0.0031064  0.1616501  -0.019 0.984668
weather_conditionlight rain. broken clouds         0.1977659  0.4416239   0.448 0.654290
weather_conditionlight rain. cloudy                0.4513946  0.2579836   1.750 0.080180 .
weather_conditionlight rain. fog                   0.0970836  0.0519406   1.869 0.061615 .
weather_conditionlight rain. low clouds           -0.2918572  0.3150567  -0.926 0.354264
weather_conditionlight rain. more clouds than sun -0.1098211  0.1584468  -0.693 0.488246
weather_conditionlight rain. mostly cloudy         0.0775449  0.0784212   0.989 0.322757
weather_conditionlight rain. overcast             -0.0781028  0.0699704  -1.116 0.264334
weather_conditionlight rain. partly cloudy         0.6243700  0.2572833   2.427 0.015239 *
weather_conditionlight rain. partly sunny          0.0888941  0.2084898   0.426 0.669840
weather_conditionlight rain. passing clouds       -0.0703732  0.1562491  -0.450 0.652432
weather_conditionlight rain. scattered clouds     -0.1700268  0.3826489  -0.444 0.656799
weather_conditionlight snow. cloudy               -0.1214897  0.2335288  -0.520 0.602904
weather_conditionlight snow. fog                  -0.5789459  0.2572564  -2.250 0.024427 *
weather_conditionlight snow. ice fog              -0.1809555  0.0795209  -2.276 0.022879 *
weather_conditionlight snow. mostly cloudy        -0.3248537  0.2915139  -1.114 0.265131
weather_conditionlight snow. overcast              0.0110393  0.1851073   0.060 0.952445
weather_conditionlow clouds                        0.0181226  0.0554784   0.327 0.743927
weather_conditionmild                              0.3334798  0.3864806   0.863 0.388220
weather_conditionmore clouds than sun              0.0121383  0.0652591   0.186 0.852444
weather_conditionmostly cloudy                     0.1012101  0.0476836   2.123 0.033801 *
weather_conditionovercast                          0.0728906  0.0451256   1.615 0.106260
weather_conditionpartly cloudy                     0.0089713  0.0842063   0.107 0.915155
weather_conditionpartly sunny                      0.0716675  0.0481830   1.487 0.136919
weather_conditionpassing clouds                    0.0798374  0.0455940   1.751 0.079947 .
weather_conditionrain. fog                         0.0256017  0.0680090   0.376 0.706588
weather_conditionrain. overcast                    0.1326477  0.2918901   0.454 0.649513
weather_conditionscattered clouds                  0.0540865  0.0508037   1.065 0.287057
```

Figure 13: Linear Model 2 Weather Condition Summary

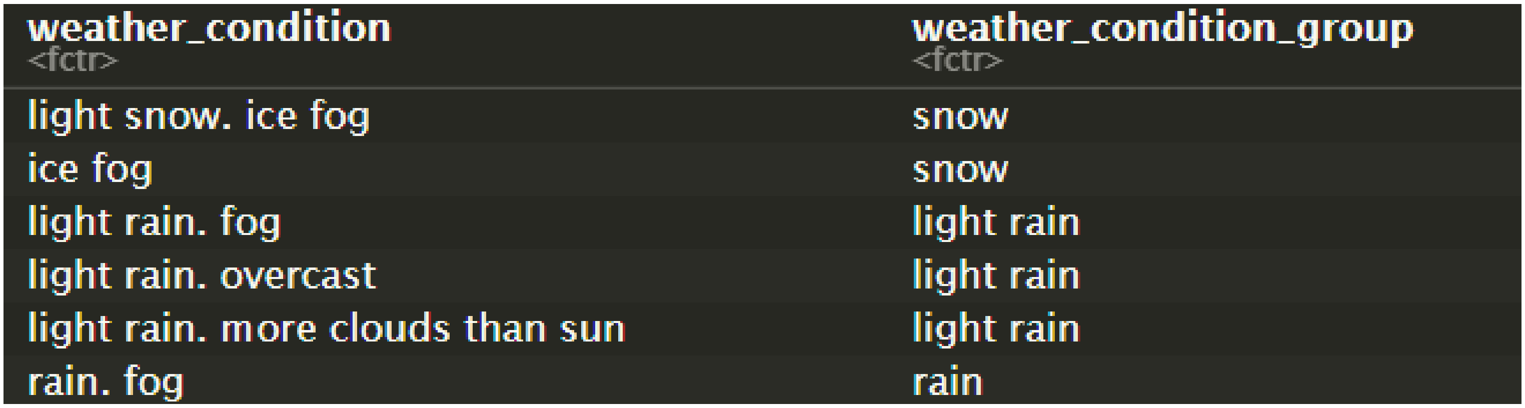
However, this study[3] which analyses the relationship between rainfall and demand for taxi concluded that demand for taxi increases as rainfall level increases, especially during rush hours. As such, I formed a hypothesis that weather_condition is not significant in this case because its factor levels are too precise to have any predicting power. There are a total of 46 "unique" weather conditions but many of them can be grouped together according to some defined criteria (in this case, I chose the "defined criteria" to be level of rain/snow).



```
 [1] overcast                      mostly cloudy                   passing clouds
 [4] clear                         sunny                           scattered clouds
 [7] light rain. overcast          light rain. more clouds than sun light rain. fog
[10] fog                           low clouds                      rain. fog
[13] partly sunny                  partly cloudy                   light snow. ice fog
[16] more clouds than sun          light snow. overcast            snow. overcast
[19] snow. ice fog                 cloudy                          rain. overcast
[22] ice fog                       light rain. mostly cloudy       snow. mostly cloudy
[25] broken clouds                 heavy snow. ice fog             light snow. cloudy
[28] haze                          chilly                          cold
[31] light freezing rain. ice fog  snow. fog                       heavy snow. fog
[34] light snow. fog               heavy rain. fog                 light rain. partly sunny
[37] light rain. low clouds        light rain. passing clouds      light rain. cloudy
[40] light rain. partly cloudy     heavy rain. partly sunny        light rain. scattered clouds
[43] mild                          light rain. broken clouds       thunderstorms. cloudy
[46] light snow. mostly cloudy
46 Levels: broken clouds chilly clear cloudy cold fog haze heavy rain. fog heavy rain. partly sunny ... thunderstorms. cloudy
```

Figure 14: All types of Weather Conditions

As such, I divided weather conditions into 7 groups (heavy snow, snow, light snow, heavy rain, rain, light rain, none). For example, all types containing keyword "light rain" can be grouped into a factor called "light rain". "Thunderstorms" can belong to the "heavy rain" group. "Ice fog" can belong to the "snow" group.

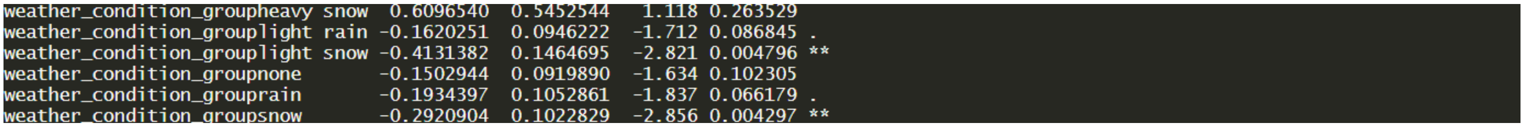


weather_condition <fctr>	weather_condition_group <fctr>
light snow. ice fog	snow
ice fog	snow
light rain. fog	light rain
light rain. overcast	light rain
light rain. more clouds than sun	light rain
rain. fog	rain

Figure 15: Sample of Weather Condition Group

Next, I re-fitted the linear model (Linear Model 3) with replacing weather_condition with weather_condition_group.

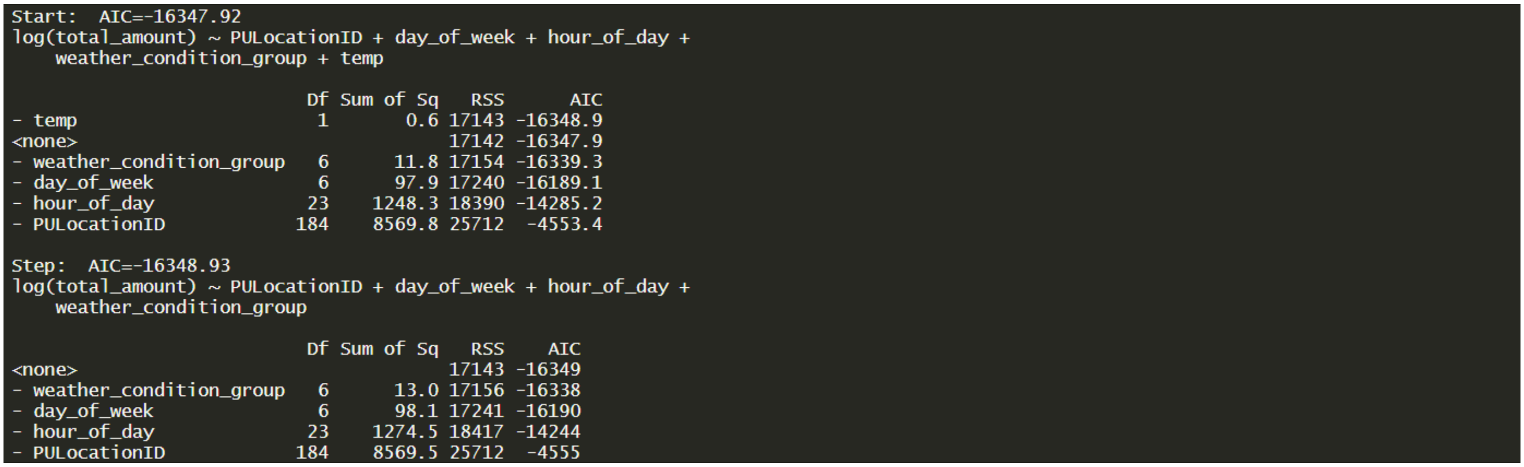
The summary shows that condition 4 and 6 (light snow and snow) are two stars significance, indicating high predicting power. This could be a better model as compared to previous model where most weather conditions are insignificant.



```
weather_condition_groupheavy snow  0.6096540  0.5452544   1.118 0.263529
weather_condition_grouplight rain -0.1620251  0.0946222  -1.712 0.086845 .
weather_condition_grouplight snow -0.4131382  0.1464695  -2.821 0.004796 **
weather_condition_groupnone       -0.1502944  0.0919890  -1.634 0.102305
weather_condition_grouprain       -0.1934397  0.1052861  -1.837 0.066179 .
weather_condition_groupsnow       -0.2920904  0.1022829  -2.856 0.004297 **
```

Figure 16: Linear Model 3 Summary Weather Condition

In addition, the linear model 3 produces an AIC of 68790.39, which is already lower than model 2 after stepwise selection. With stepwise selection, AIC drops to 68789.38, which is even better. But interestingly, unlike stepwise selection from model 2, this selection dropped temperature instead of weather_condition_group. This could suggest that weather_condition_group has became a better indicator than temperature.



```
Start:  AIC=-16347.92
log(total_amount) ~ PULocationID + day_of_week + hour_of_day + 
    weather_condition_group + temp

                          Df Sum of Sq   RSS      AIC
- temp                     1       0.6 17143 -16348.9
<none>                                 17142 -16347.9
- weather_condition_group  6      11.8 17154 -16339.3
- day_of_week              6      97.9 17240 -16189.1
- hour_of_day             23    1248.3 18390 -14285.2
- PULocationID           184    8569.8 25712  -4553.4

Step:  AIC=-16348.93
log(total_amount) ~ PULocationID + day_of_week + hour_of_day + 
    weather_condition_group

                          Df Sum of Sq   RSS    AIC
<none>                                 17143 -16349
- weather_condition_group  6      13.0 17156 -16338
- day_of_week              6      98.1 17241 -16190
- hour_of_day             23    1274.5 18417 -14244
- PULocationID           184    8569.5 25712  -4555
```

Figure 17: Stepwise 2

5.4 Interaction between temperature and weather condition

This box plot of temperature against weather condition shows that temperature is lower when weather condition is snowy and higher when it is rainy or none. This implied there might be interaction between temperature and weather condition.

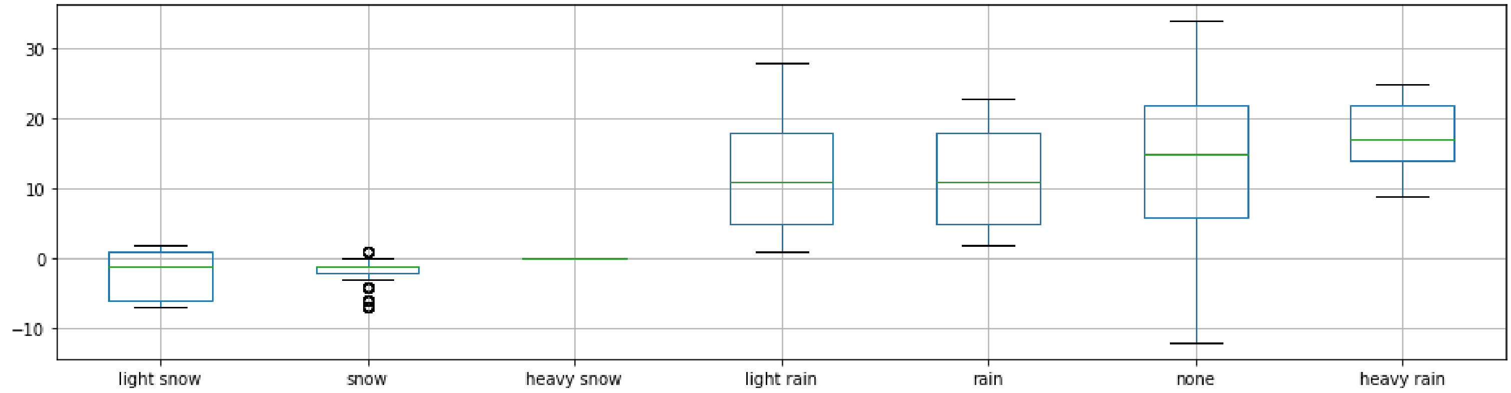
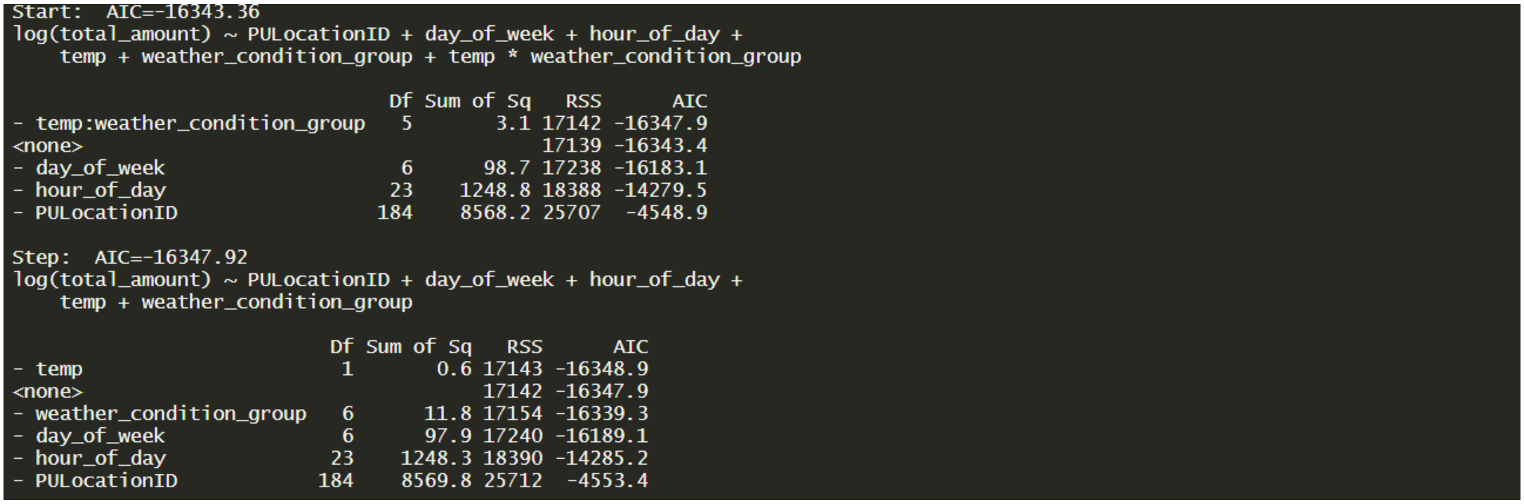


Figure 18: Temperature vs Weather Condition

As such, I fitted linear model 4 with the additional interaction term temp*weather_condition_group. However, the interaction term turns out to be insignificant as stepwise selection drops this term immediately at the start.

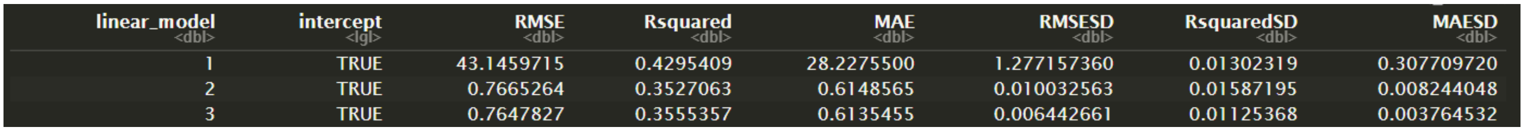


```
Start:  AIC=-16343.36
log(total_amount) ~ PULocationID + day_of_week + hour_of_day + 
    temp + weather_condition_group + temp * weather_condition_group

                               Df Sum of Sq   RSS      AIC
- temp:weather_condition_group  5       3.1 17142 -16347.9
<none>                                      17139 -16343.4
- day_of_week                   6      98.7 17238 -16183.1
- hour_of_day                  23    1248.8 18388 -14279.5
- PULocationID                184    8568.2 25707  -4548.9

Step:  AIC=-16347.92
log(total_amount) ~ PULocationID + day_of_week + hour_of_day + 
    temp + weather_condition_group

                          Df Sum of Sq   RSS      AIC
- temp                     1       0.6 17143 -16348.9
<none>                                 17142 -16347.9
- weather_condition_group  6      11.8 17154 -16339.3
- day_of_week              6      97.9 17240 -16189.1
- hour_of_day             23    1248.3 18390 -14285.2
- PULocationID           184    8569.8 25712  -4553.4
```

Figure 19: Stepwise 3

5.5 Cross Validation

Next, I performed 5-fold cross validation on linear model 1, 2 and 3 (2 and 3 are models after stepwise selection) with Caret and obtained the prediction result.



linear_model	intercept	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
<dbl>	<lgl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	TRUE	43.1459715	0.4295409	28.2275500	1.277157360	0.01302319	0.307709720
2	TRUE	0.7665264	0.3527063	0.6148565	0.010032563	0.01587195	0.008244048
3	TRUE	0.7647827	0.3555357	0.6135455	0.006442661	0.01125368	0.003764532

Figure 20: Cross Validation Results

Model 2 and 3 have significantly smaller Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) than model 1, indicating that log transformation is very effective. Model 3 has smaller RMSE and MAE than model 2, indicating that grouping weather condition improves accuracy. Also, model 3 has a larger Rsquared than model 2, indicating that model 3 fits the data and explaining the variability of data better than model 2.

6 Evaluation

In this section, I ranked predictors based on their significance and correlation with response. Also, I provided recommendation for taxi company in task-assignment according to predictor variables.

6.1 Ranking Significance of Predictors

From model selection, our final model includes PULocationID (Factor), day_of__the_week (Factor), hour_of_the_day (Factor), weather_condition_group (Factor) as predictors and log(total_amount) as response. I first consider p-value of each predictor in the presence of other predictors. Note that this method only allows us to compare factor levels not among factors themselves. I visualize pvalues of taxi zones.

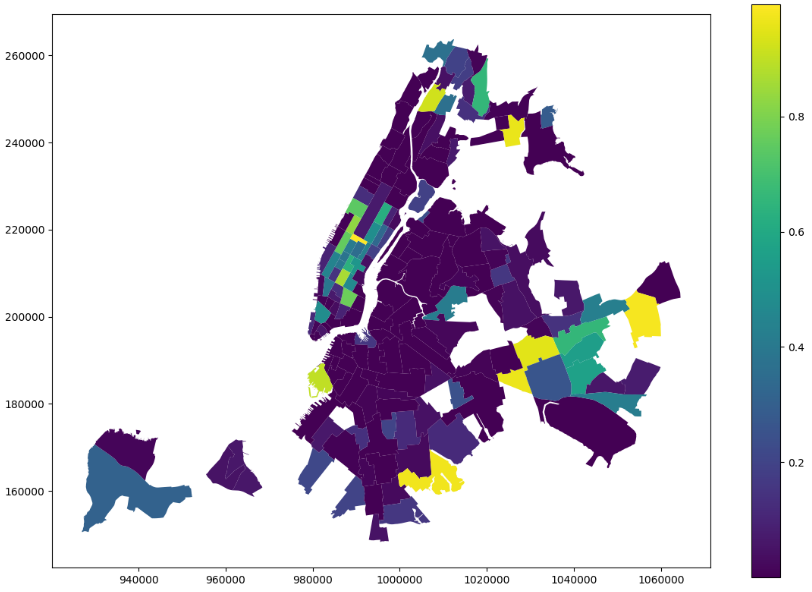
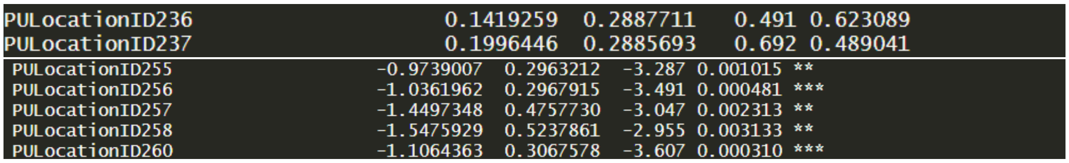


Figure 21: pvalues by taxi zone

I observed that p-values of zones at central manhattan regions and southern Queens are significantly higher (greater than 0.4) than other regions, indicating that these zones have less impact on response. However, most of these zones have positive estimated coefficient, indicating an association with increase in hourly total amount. On the other hand, most zones in Brooklyn and northern Queens have p-value less than 0.1, indicating higher influence on varying hourly total earning. However, these zones have negative estimated coefficient, indicating an association with decrease in total amount, therefore higher probability of earning less.

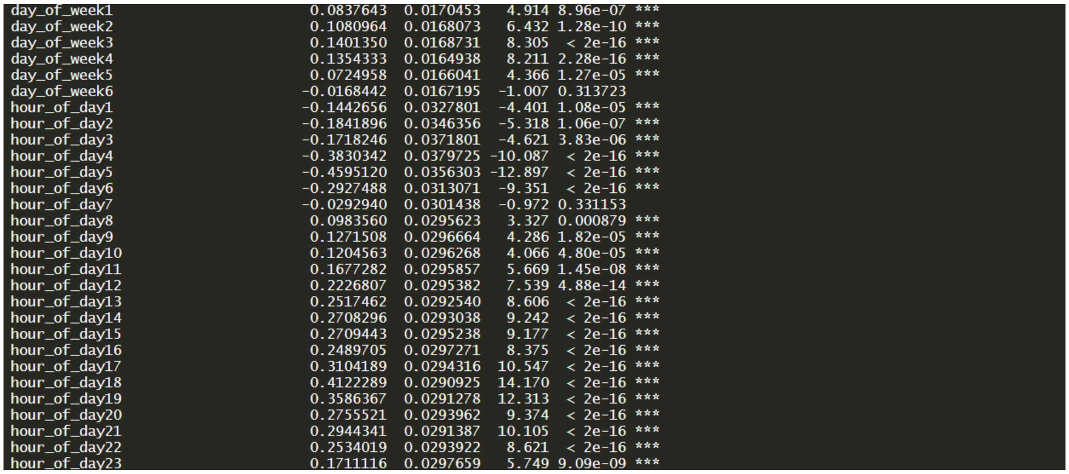


```
PULocationID236                   0.1419259  0.2887711   0.491 0.623089
PULocationID237                   0.1996446  0.2885693   0.692 0.489041
```

```
PULocationID255               -0.9739007  0.2963212  -3.287 0.001015 **
PULocationID256               -1.0361962  0.2967915  -3.491 0.000481 ***
PULocationID257               -1.4497348  0.4757730  -3.047 0.002313 **
PULocationID258               -1.5475929  0.5237861  -2.955 0.003133 **
PULocationID260               -1.1064363  0.3067578  -3.607 0.000310 ***
```

Figure 22: Selected zones in Central Manhattan vs Selected zones in Brooklyn and Queens

On the other hand, hour of the day and day of the week levels are highly significant (most of which are 3 stars), indicating that they have high association with hourly total earning.

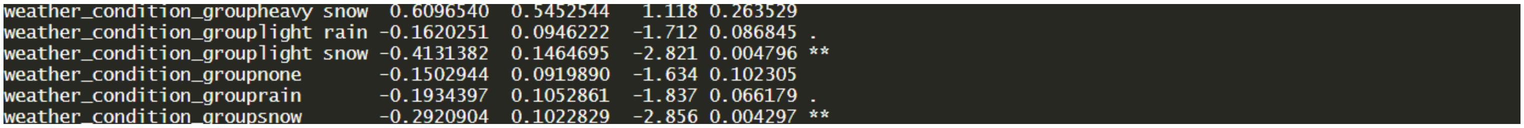


```
day_of_week1                      0.0837643  0.0170453   4.914 8.96e-07 ***
day_of_week2                      0.1080964  0.0168073   6.432 1.28e-10 ***
day_of_week3                      0.1401350  0.0168731   8.305  < 2e-16 ***
day_of_week4                      0.1354333  0.0164938   8.211 2.28e-16 ***
day_of_week5                      0.0724958  0.0166041   4.366 1.27e-05 ***
day_of_week6                     -0.0168442  0.0167195  -1.007 0.313723
hour_of_day1                     -0.1442656  0.0327801  -4.401 1.08e-05 ***
hour_of_day2                     -0.1841896  0.0346356  -5.318 1.06e-07 ***
hour_of_day3                     -0.1718246  0.0371801  -4.621 3.83e-06 ***
hour_of_day4                     -0.3830342  0.0379725 -10.087  < 2e-16 ***
hour_of_day5                     -0.4595120  0.0356303 -12.897  < 2e-16 ***
hour_of_day6                     -0.2927488  0.0313071  -9.351  < 2e-16 ***
hour_of_day7                     -0.0292940  0.0301438  -0.972 0.331153
hour_of_day8                      0.0983560  0.0295623   3.327 0.000879 ***
hour_of_day9                      0.1271508  0.0296664   4.286 1.82e-05 ***
hour_of_day10                     0.1204563  0.0296268   4.066 4.80e-05 ***
hour_of_day11                     0.1677282  0.0295857   5.669 1.45e-08 ***
hour_of_day12                     0.2226807  0.0295382   7.539 4.88e-14 ***
hour_of_day13                     0.2517462  0.0292540   8.606  < 2e-16 ***
hour_of_day14                     0.2708296  0.0293038   9.242  < 2e-16 ***
hour_of_day15                     0.2709443  0.0295238   9.177  < 2e-16 ***
hour_of_day16                     0.2489705  0.0297271   8.375  < 2e-16 ***
hour_of_day17                     0.3104189  0.0294316  10.547  < 2e-16 ***
hour_of_day18                     0.4122289  0.0290925  14.170  < 2e-16 ***
hour_of_day19                     0.3586367  0.0291278  12.313  < 2e-16 ***
hour_of_day20                     0.2755521  0.0293962   9.374  < 2e-16 ***
hour_of_day21                     0.2944341  0.0291387  10.105  < 2e-16 ***
hour_of_day22                     0.2534019  0.0293922   8.621  < 2e-16 ***
hour_of_day23                     0.1711116  0.0297659   5.749 9.09e-09 ***
```

Figure 23: pvalues

From Monday and Saturday, coefficients ar positive hence driving would more likely to be profitable. On the other hand, driving in Sunday may earn less as there is a slight negative effect on average earning. As for hour of the day, it is better to drive from 08:00 to 23:00 as p values during this period is small and coefficient is positive. However, driving from 1am to 6am would more likely earn less again because of negative coefficient.

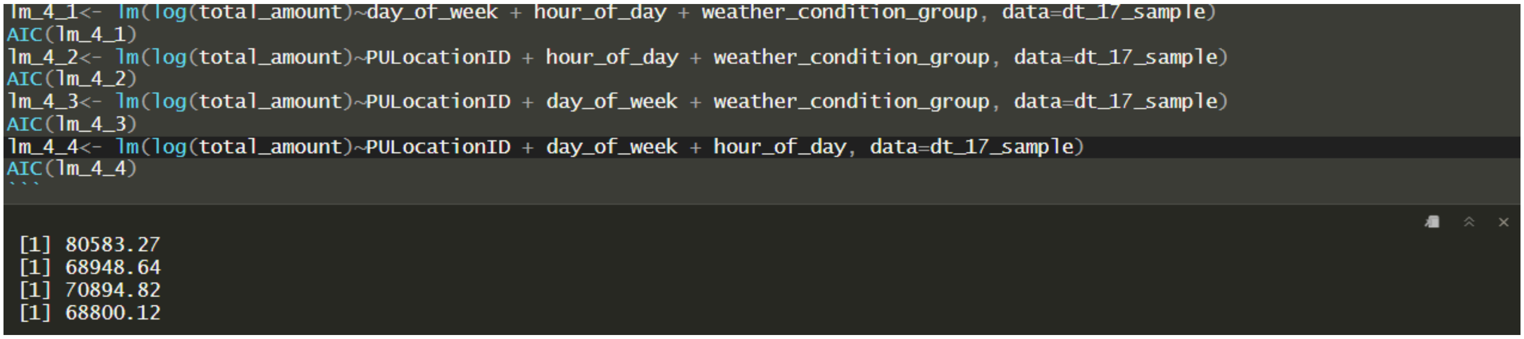
Weather condition group, on the other hand, have fewer significant levels. Also, only heavy snow has positive association with total amount whereas light snow and snowy conditions have large negative coefficient. This suggests driving under heavy snow condition is more profitable than other conditions.



```
weather_condition_groupheavy snow  0.6096540  0.5452544   1.118 0.263529
weather_condition_grouplight rain -0.1620251  0.0946222  -1.712 0.086845 .
weather_condition_grouplight snow -0.4131382  0.1464695  -2.821 0.004796 **
weather_condition_groupnone       -0.1502944  0.0919890  -1.634 0.102305
weather_condition_grouprain       -0.1934397  0.1052861  -1.837 0.066179 .
weather_condition_groupsnow       -0.2920904  0.1022829  -2.856 0.004297 **
```

Figure 24: Linear Model 3 Summary Weather Condition

Next, I compared significance among predictors by fitting 4 different models with each predictor removed and compared their AICs.



~~~
lm_4_1<- lm(log(total_amount)~day_of_week + hour_of_day + weather_condition_group, data=dt_17_sample)
AIC(lm_4_1)
lm_4_2<- lm(log(total_amount)~PULocationID + hour_of_day + weather_condition_group, data=dt_17_sample)
AIC(lm_4_2)
lm_4_3<- lm(log(total_amount)~PULocationID + day_of_week + weather_condition_group, data=dt_17_sample)
AIC(lm_4_3)
lm_4_4<- lm(log(total_amount)~PULocationID + day_of_week + hour_of_day, data=dt_17_sample)
AIC(lm_4_4)
```
~~~

```
[1] 80583.27
[1] 68948.64
[1] 70894.82
[1] 68800.12
```

Figure 25: AICs

The result shows that AIC is the highest when PULocationID is removed, indicating that it has the highest influence on response which is followed by hour of the day, day of the week and weather condition group. This result concludes that the most important factor affecting profitability is location followed by period of the day. On the other hand, day of the weather and weather condition are influential but need not to be prioritized.

7 Conclusion

In this study, I explored the linear relation between pick-up taxi zone, day of the week, weather condition and temperature at a hour of the day, with total earning in this hour. I discovered that taxi zone, day of the week, and hour of day are closely related to log of total earning whereas temperature and weather conditions are less but still significant. Dividing weather conditions into group improves their explanatory power and improves overall model fitting. In terms of strategic assignment of tasks to taxis, it may be more profitable to allocate more taxis in central manhattan instead of Brooklyn and northern Queens. Driving during the day would be better than after midnight and driving from Monday to Saturday is better than Sunday. Although weather conditions are less significant, it would be more profitable driving under heavy snow condition as opposed to light snow or snowy conditions. As for future work, I may consider different other transformation to models, exploring non-parametric models and experiment with other external factors like major events.

References

[1] N. L. Pesce, “This chart shows how uber rides sped past nyc yellow cabs in just six years,” Aug 2019.

[2] Rong-Gang, Brady, and Mark, “The interdependence between rainfall and temperature: Copula analyses,” Nov 2012.

[3] D. Chen, Y. Zhang, L. Gao, N. Geng, and X. Li, “The impact of rainfall on the temporal and spatial distribution of taxi passengers,” Sep 2017.