Final Project: Uber Pickups in New York City

Varun Prasad

December 10, 2019

Summary

This report investigates the factors that affect the number of hourly Uber pickups across the five boroughs of New York City: The Bronx, Brooklyn, Manhattan, Queens, and Staten Island. The data analyzed contained information about the number of pickups per hour in each borough in addition to weather conditions and whether the day was a holiday. The final valid model was a hierarchical negative binomial regression with borough treated as the random intercept. The results showed that higher values of wind speed, temperature, and sea level pressure increased Uber pickups while higher values of 24-hour precipitation levels, snow depth, and visibility reduced Uber pickups. Since the dataset only consisted of the first half of 2015, further testing should be done on a full year for a more complete and valid model. Still, this model does provide insight into the factors that influence Uber pickups, and future work can incorporate time series analysis, particularly with weather forecasting, to predict future Uber pickups around New York City.

Introduction

New York City is divided into the five following boroughs: The Bronx, Brooklyn, Manhattan, Queens, and Staten Island. Each borough is coextensive with its respective county and has its own governmental administration. The most familiar and densely populated of these boroughs is Manhattan, which contains many of New York City's most famous cultural and economic centers. For many years, people traveled through these boroughs using the iconic yellow taxi cab, but recently, Uber has become much more commonly used by tourists and residents of New York City. The analysis presented in this report investigates the factors that affect the number of Uber pickups per hour in each of the five boroughs. Specific questions that will be addressed include the following:

1. Do holidays significantly affect the number of Uber pickups compared to normal days?
2. Do increased levels of rain and snow decrease Uber pickups?
3. Do conditions such as low visibility and low temperature lead to an increase in pickups?

Data

The dataset used for this analysis is a subset of FiveThirtyEight's *Uber Pickups in New York City* dataset, with resulting dates of pickups ranging from January 1, 2015 to June 30, 2015. A user on Kaggle named Yannis Pappas has merged this data with weather data from the National Centers for Environmental Information and with location ID data from FiveThirtyEight to map Uber pickups for each borough in NYC at specific times and with corresponding weather conditions. The dataset can be found here.

The dataset contains 29,101 observations of pickups. The number of pickups for each borough per hour has been recorded, along with variables such as wind speed (mph), visibility (miles to nearest tenth), temperature (^{o}F), dew point (^{o}F), sea level pressure (mbar), precipitation (in inches at 1, 6, and 24 hours), snow depth (inches), and whether or not the day was a holiday. The boroughs include the main five, Newark Aiport (EWR), and "NA", the last of which indicates a pickup outside a borough or one that has an unidentified mapping. The "EWR" borough had a maximum of two counts per hour while the "NA" borough had single

digit counts as well as fewer overall observations compared to the other boroughs. Thus, all observations corresponding to the "EWR" and "NA" boroughs were removed, thereby ensuring focus on just the main five boroughs. The final dataset had 21715 observations and 13 variables. The code book for this dataset can be found in Appendix 1.1.

Exploratory data analysis first focused on the distribution of Uber pickups in each of the five boroughs, which is shown in the following boxplot:

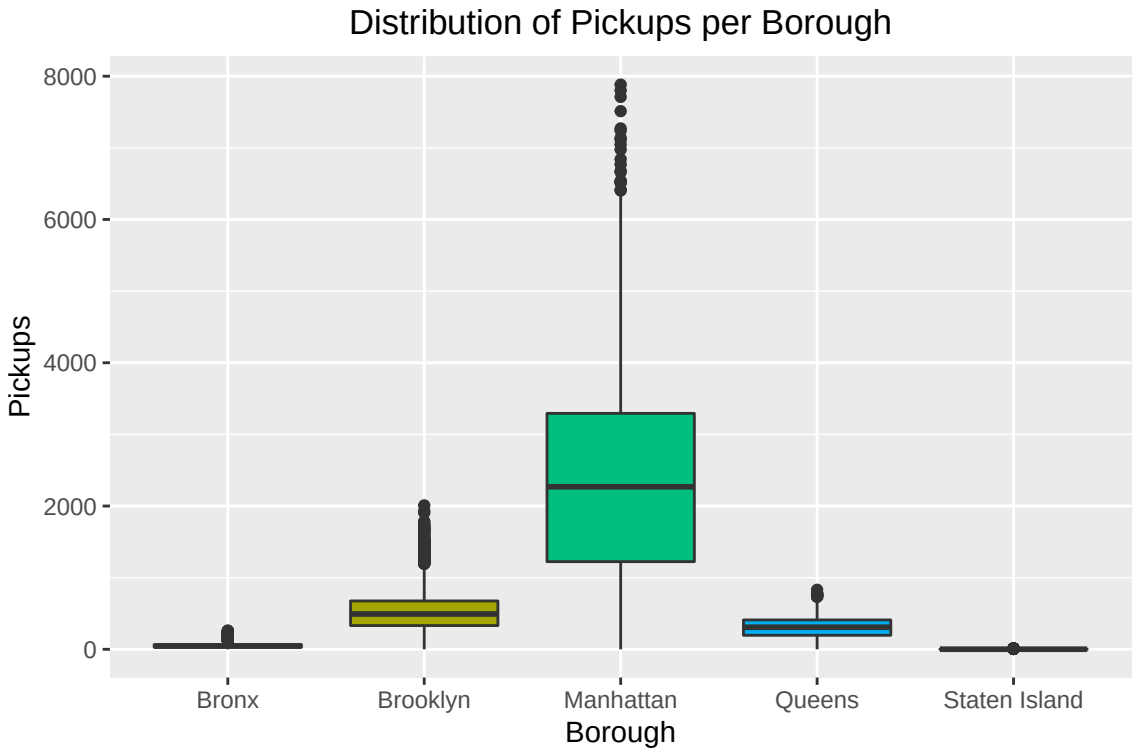


Figure 1: Distribution of Uber Pickups in Each Borough

The boxplot shows that Manhattan has by far the largest average number of pickups per hour amongst the five boroughs, followed by Brooklyn, Queens, Bronx, and State Island. This plot suggests that a hierarchical model should ultimately be used to account for these vastly different baseline pickup counts between the boroughs.

Additional EDA focused on the predictor variables. A correlation matrix, shown in Appendix 1.2, revealed a value of 0.89 between temperature and dew point. This result also held after mean centering the variables. Thus, dew point was removed from further analysis. There were seven holidays in the dataset, the most notable of which are New Years' Day (Jan 1), Martin Luther King Jr.'s birthday (Jan 19), Memorial Day (May 25), and Presidents' Day (Feb 16). The boxplot below shows the distribution of pickups in each borough on both holidays and non-holidays. Holidays do not appear to significantly affect the average number of pickups, though a higher maximum of pickups is observed on non-holidays.

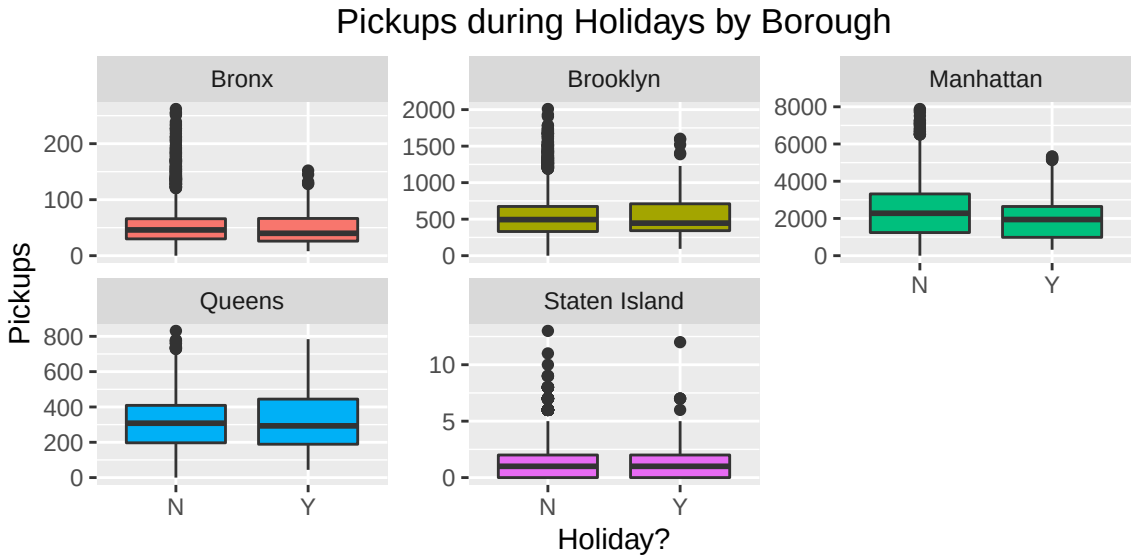


Figure 2: Pickups During Holidays by Borough

Plots of the continuous variables against pickups by borough showed some notable trends using a linear fit. In particular, pickups appeared to increase with temperature and decrease with higher levels of 6-hour precipitation, 24-hour precipitation (shown in Figure 3), and sea level pressure. Snow depth, 1-hour precipitation, visibility, and wind speed did not appear to significantly affect pickup counts. Given their geographical proximity, the boroughs all had very similar weather conditions and thus very similar trends. Additional plots of the predictor variables against pickups by borough are shown in Appendix 1.3.

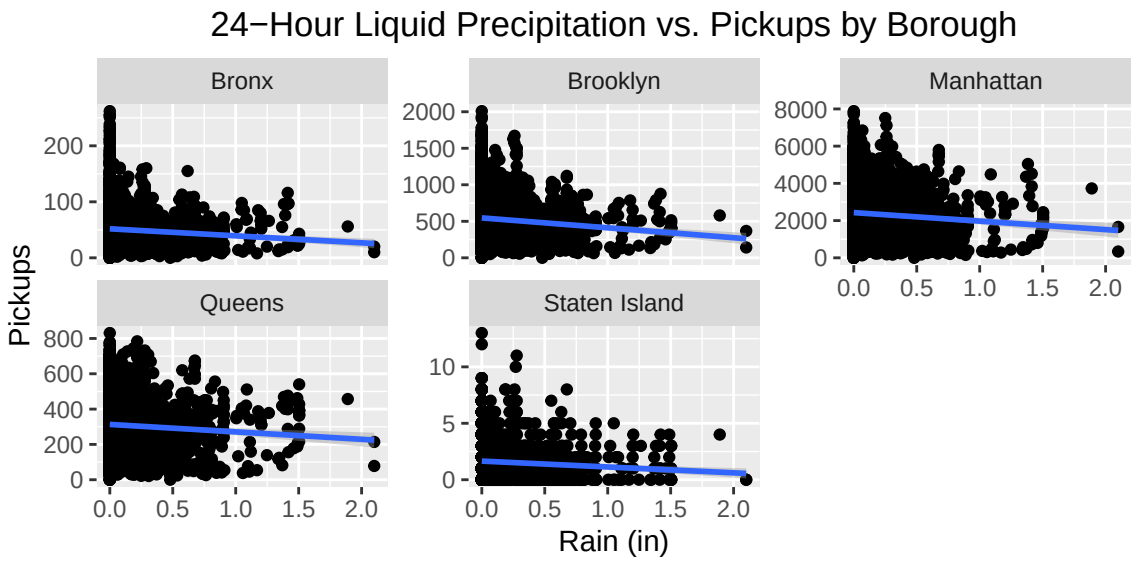


Figure 3: 24-Hour Precipitation (in) vs. Pickups by Borough

Finally, several interactions were explored. Plots of temperature versus snow depth showed that after a temperature of approximately 60^{o}F, snow depth was constant at 0 inches. Furthermore, snow depth tended to decrease as temperature increased, as expected. However, there were instances of high snow depth at temperatures above the melting point, indicating a recent period of warmth after a snowstorm. In addition,

as temperature increased, sea level pressure tended to decrease, which is also expected. Interactions will be further tested during model building, and plots can be seen in Appendix 1.4.

Model/Results

Prior to model building, all continuous variables were mean-centered. Since the response variable of Uber pickups is count data, the first model that should be tested is a Poisson regression, specifically a hierarchical Poisson regression using borough as the varying-intercept. This type of model has the following general formula:

$$log(\lambda_{ij}) = \beta_{0j} + \beta_1 x_{ij} + ... + \beta_p x_{ij};\ i = 1, ..., n_j;\ j = 1, ..., J$$

$$\beta_{0j} \sim (\beta_0, \tau^2)$$

where λ is the Poisson distribution parameter, i indexes observations, j indexes groups, and τ^2 is the between group (intercept) variance.

An important assumption of the Poisson distribution is that its mean is equal to its variance. Overdispersion occurs when the variance is greater than the mean, thus invalidating the regular Poisson model. A couple Poisson models were tested, both hierarchical and with borough as a factor. In both cases, it was determined that there was significant overdispersion. This was confirmed visually by observing the very high residuals on the residual vs. fitted plot, by using the quasipoisson model instead to account for overdispersion, and by using a dispersion test from the AER package. The latter two methods assume the variance increases linearly with the mean and revealed an overdispersion parameter of approximately 208. As a result, the most appropriate model for this data is the multilevel negative binomial regression model, which contains an overdispersion parameter and can model overdispersed count data. This model was created using the glmmTMB package and more appropriately assumes variance increases quadratically with the mean.

The final model was fit with the following mean-centered parameters: wind speed, visibility, temperature, sea level pressure, 6-hour precipitation, 24-hour precipitation, snow depth, and an interaction between temperature and snow depth. A non-hierarchical model, with borough as a categorical variable, was originally developed through a backward selection process using AIC. Both holiday and 1-hour precipitation were found to be insignificant parameters. Additional interaction effects, including those between temperature and sea level pressure, 24-hour precipitation and visibility, and temperature and 24-hour precipitation were found to be insignificant. The varying intercept for borough was then applied and the model's parameters were retested with manual selection and F-tests. The final model had an intercept variance of 6.318 and an overdispersion parameter of 3. The coefficients of the model are shown in the following table:

Variable	Estimate	Exponentiated Estimate	Std. Error	z value	Pr(>|z|)
Intercept	4.760e+00	116.746	1.124e+00	4.23	2.29e-05
spd_c	1.099e-02	1.0110	1.245e-03	8.82	<2e-16
vsb_c	-6.231e-03	0.9938	1.781e-03	-3.50	0.000466
temp_c	1.003e-02	1.0101	3.020e-04	33.23	< 2e-16
slp_c	1.671e-03	1.0017	5.780e-04	2.89	0.003835
pcp06_c	8.504e-02	1.0888	4.704e-02	1.81	0.070664
pcp24_c	-2.169e-01	0.8050	1.983e-02	-10.94	< 2e-16
sd_c	-4.210e-03	0.9958	1.754e-03	-2.40	0.016347
temp_c:sd_c	-1.160e-03	0.9988	8.416e-05	-13.78	<2e-16

Table 1: Coefficients of Multilevel Negative Binomial Model

The model was validated by plotting the fitted values against the Pearson residuals and by plotting the continuous variables against the Pearson residuals. None of the plots showed any discernible pattern about 0, thus validating the model. These plots are shown in Appendix 2. The model also did not have any multicollinearity after testing the variables just prior to creating borough as a level.

From the results in Table 1, several key insights can be determined. Because the negative binomial can be linked to the logit function, the coefficients can be exponentiated to obtain more meaningful conclusions. The intercept indicates a baseline pickup count of approximately 117 when all variables are centered at their respective means. The baseline pickup count for each borough can be determined by exponentiating the respective variance and multiplying it by the intercept. The absence of holiday as a parameter indicates that holiday did not significantly affect the number of Uber pickups. Furthermore, 6-hour precipitation is also not significant. With all else constant, each unit increase in the variables of visibility, 24-hour liquid precipitation, and snow depth reduce the number of Uber pickups by 0.62%, 19.5%, and 0.42%, respectively. With all else constant, each unit increase in the variables of wind speed, temperature and sea level pressure increase Uber pickups by 1.1%, 1.01%, and 0.17%, respectively. The interaction term between temperature and snow depth indicates that at a constant temperature, each inch increase in snow depth reduces Uber pickups by 0.12%. These results are largely consistent with the exploratory data analysis.

The dotplot below further highlights the great variation in pickups between the boroughs. The model results further validate the EDA since they show Manhattan with the greatest positive intercept followed by Brooklyn, Queens, Bronx, and Staten Island.

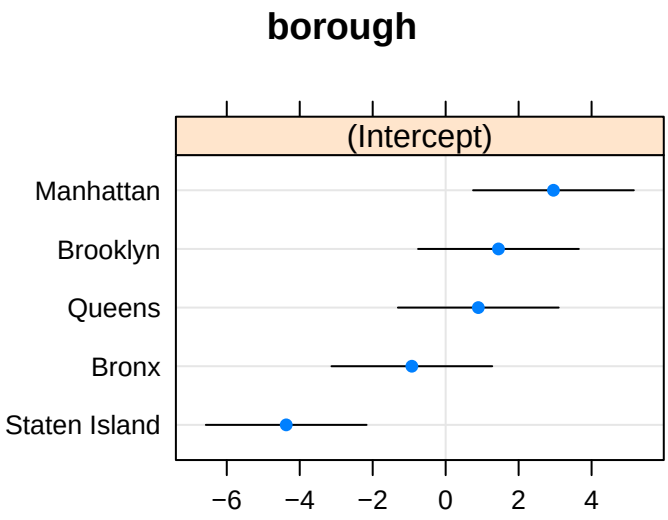


Figure 4: Dotplot of Boroughs from Multilevel Model

Conclusions

Overall, the model was valid and addressed questions about the factors that affected the number of Uber pickups across the five boroughs of New York City. Specifically, it was discovered that holidays were insignificant, that higher levels of rain and snow reduced Uber pickups, and that higher temperatures led to an increase in pickups. However, there are several limitations in this analysis. The data only contains pickups from January 1 to June 30 of 2015. Having data for the remaining half of the year, and for more years, would encompass more weather and pickup data, thereby improving the model and strengthening the conclusions. In addition, having a full year of data would include more popular holidays such as Independence Day, Thanksgiving, and Christmas. Incorporating all of these holidays could lead to a different conclusion about the effect of holidays on pickups. Still, the results from this model do provide a strong foundation for future work. Population and demographic data about Uber users could provide greater insight into what types of people use Uber and if these factors vary across boroughs. Future work could also incorporate a time series model with weather, thereby creating a joint forecast of Uber pickups from weather data.

Appendices

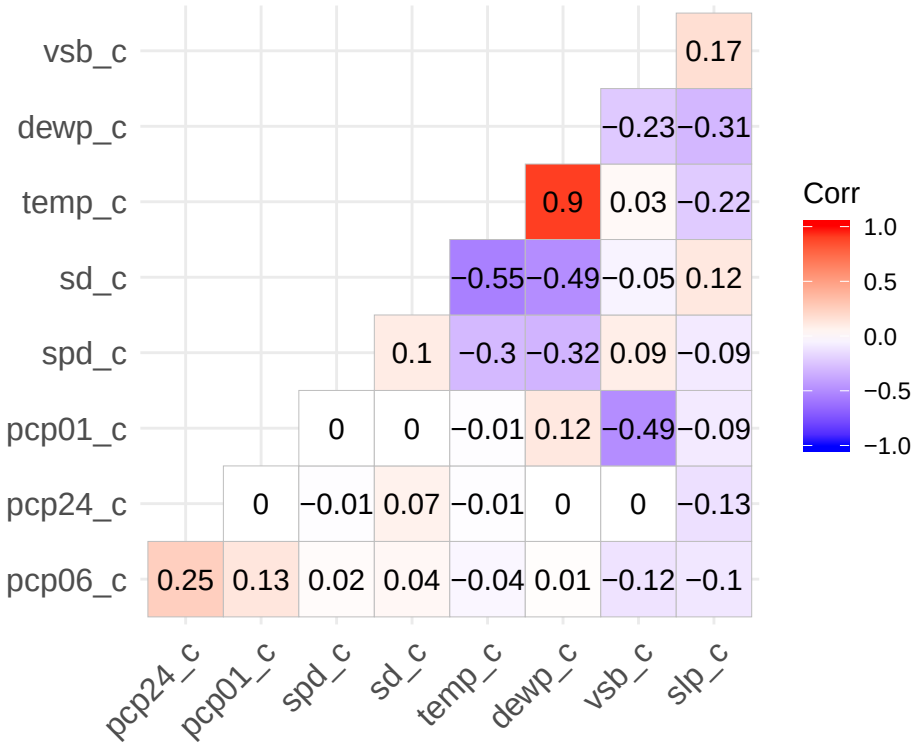
Appendix 1: EDA

Appendix 1.1: Table of Variables

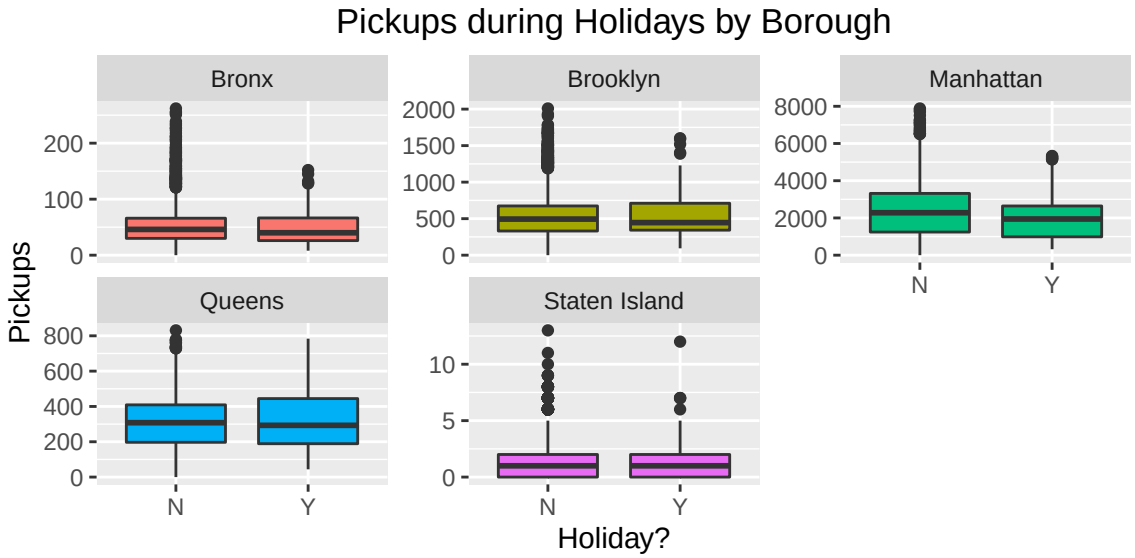
Variable	Description
pickup_dt	Time period of observation
borough	NYC's borough
pickups	Number of pickups in time period
spd	Wind speed (mph)
vsb	Visibility (miles to nearest tenth)
temp	Temperature (^{o}F)
dewp	Dew point (^{o}F)
slp	Sea level pressure (mbar)
pcp01	1-hour liquid precipitation (in)
pcp06	6-hour liquid precipitation (in)
pcp24	24-hour liquid precipitation (in)
sd	snow depth (in)
hday	Holiday (Y or N)

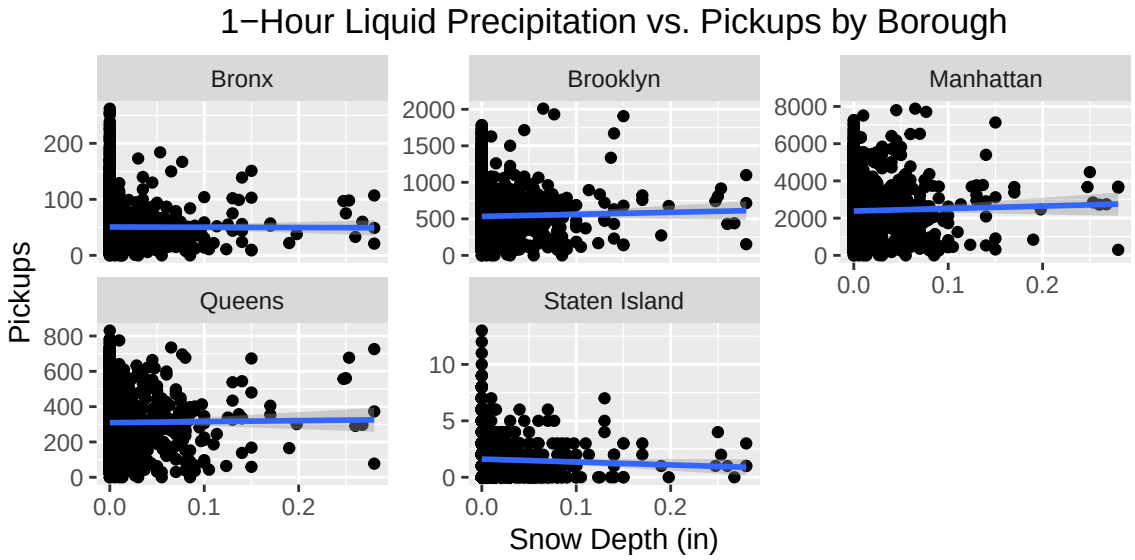
Table 2: Code Book of Dataset

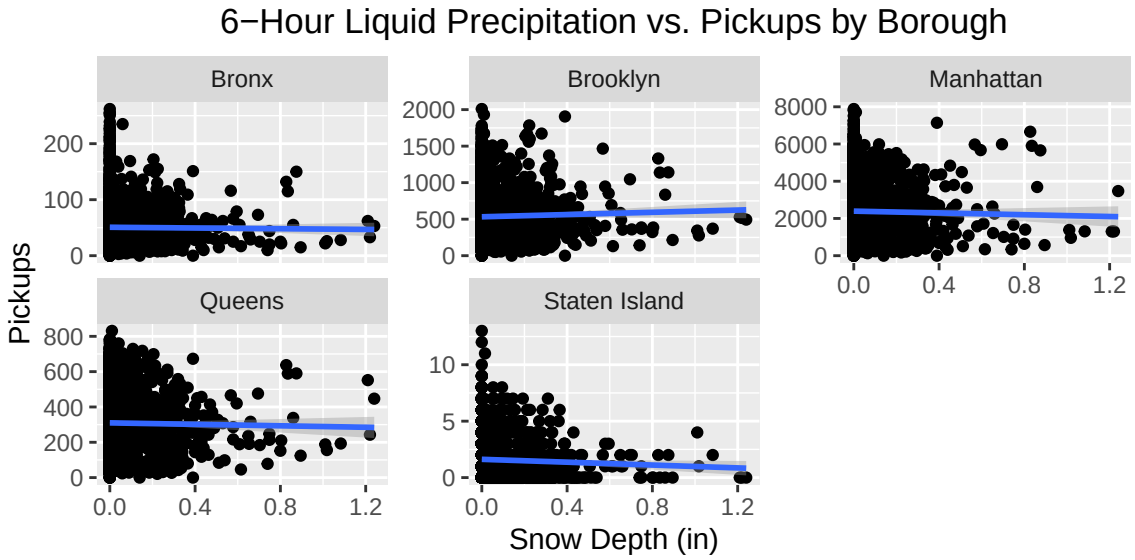
Appendix 1.2: Correlation Matrix

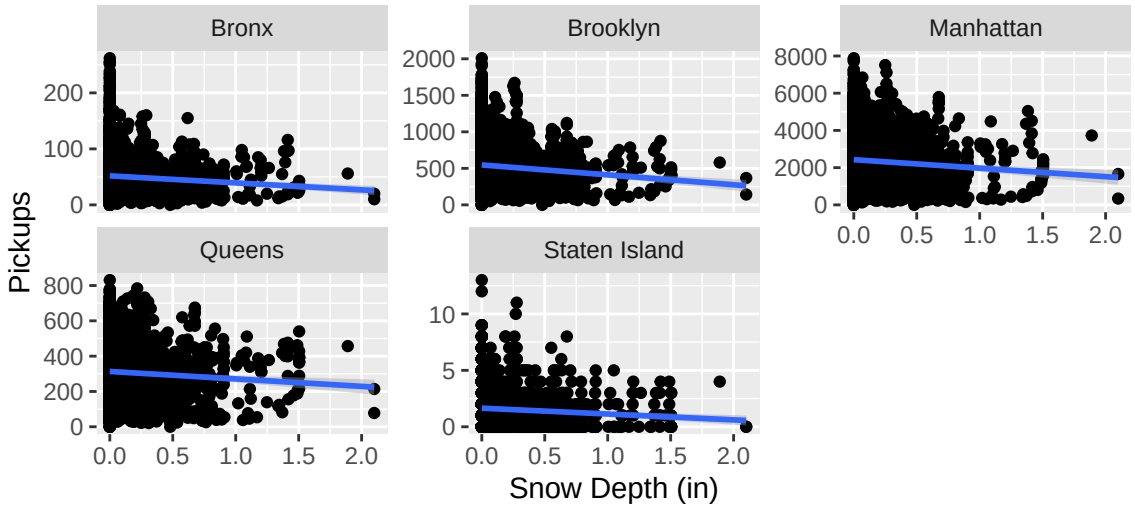


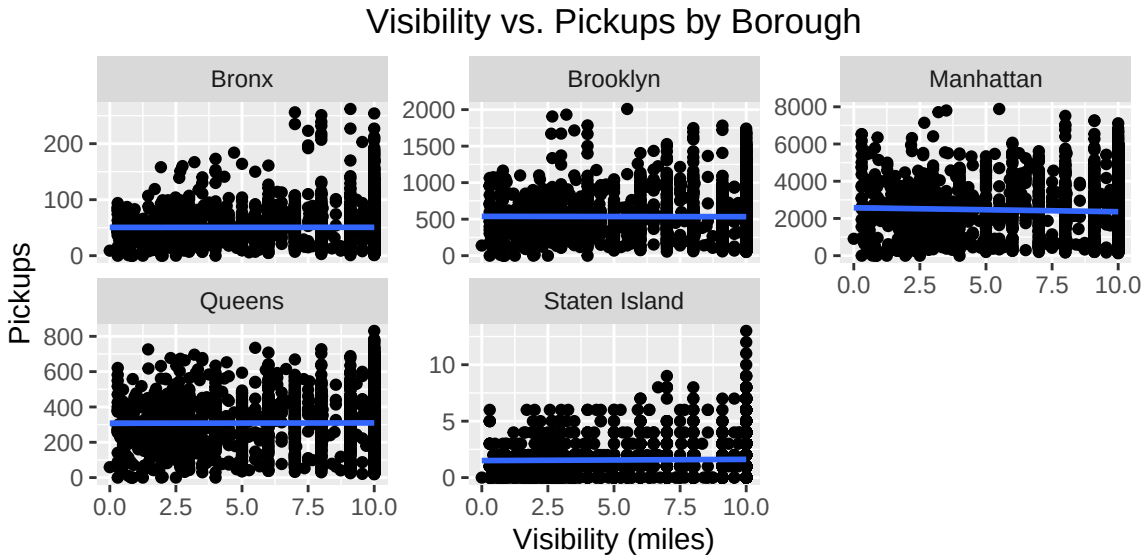
Appendix 1.3: Pickups vs. Predictors

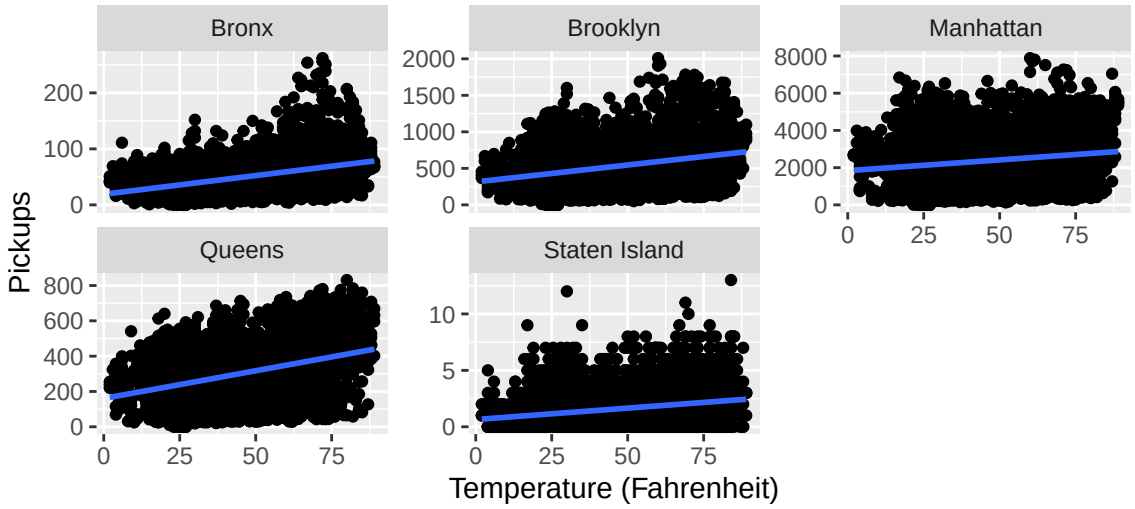


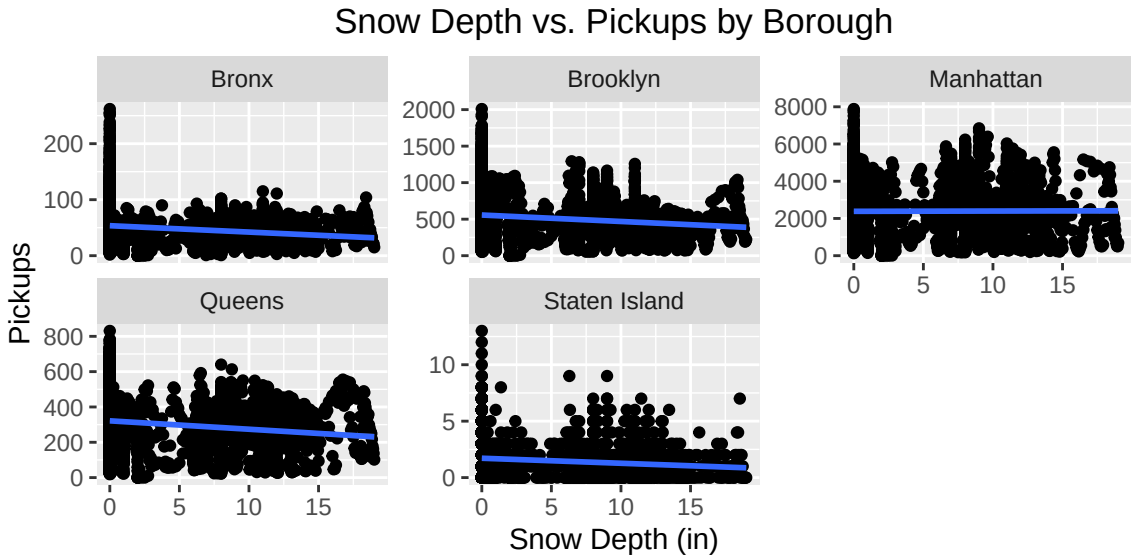


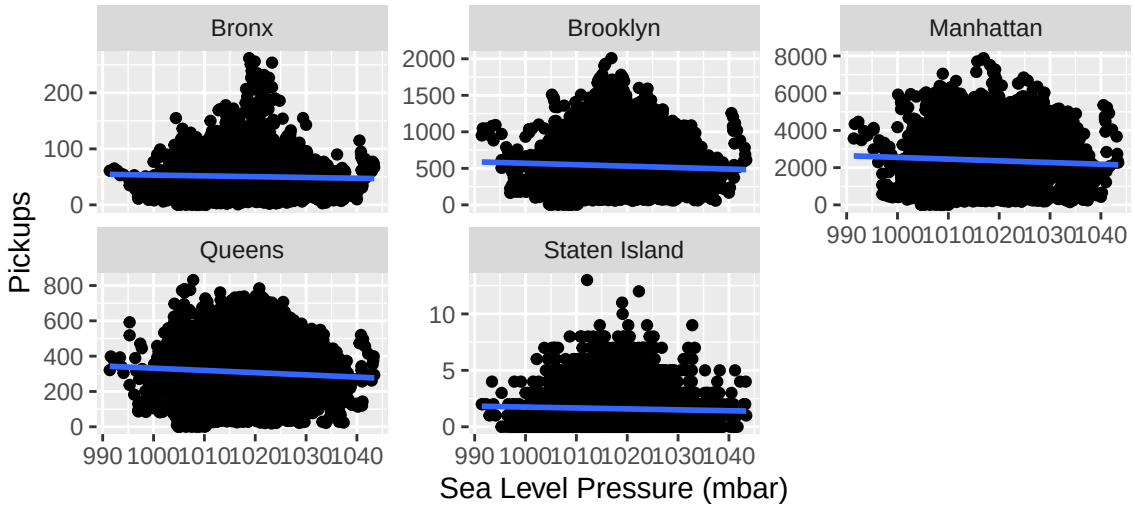

6−Hour Liquid Precipitation vs. Pickups by Borough
Bronx
Brooklyn
Manhattan
Queens
Staten Island
Pickups
Snow Depth (in)

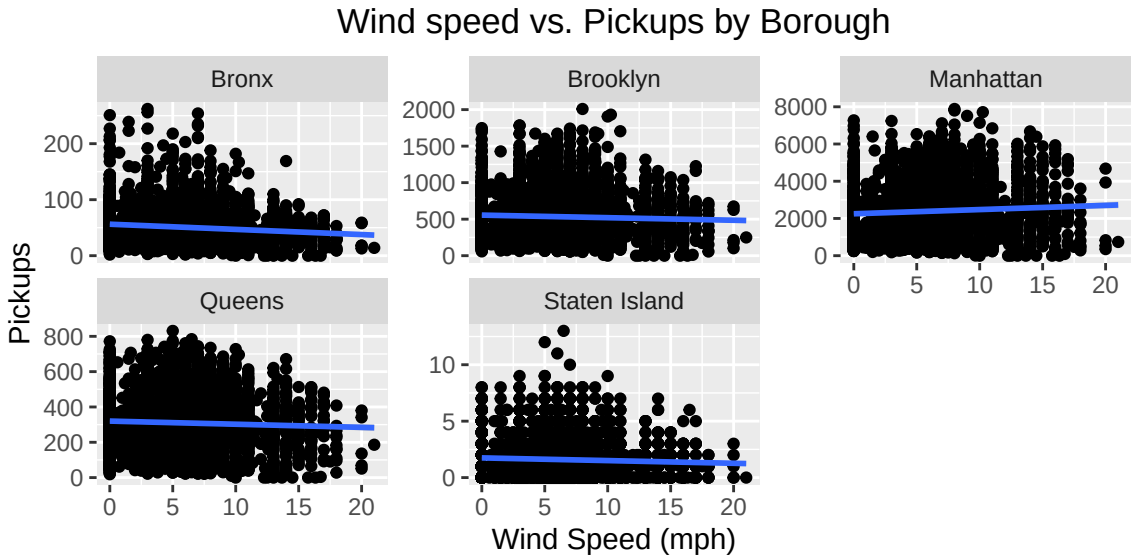

24−Hour Liquid Precipitation vs. Pickups by Borough
Bronx
Brooklyn
Manhattan
Queens
Staten Island
Pickups
Snow Depth (in)


Visibility vs. Pickups by Borough
Bronx
Brooklyn
Manhattan
Queens
Staten Island
Pickups
Visibility (miles)

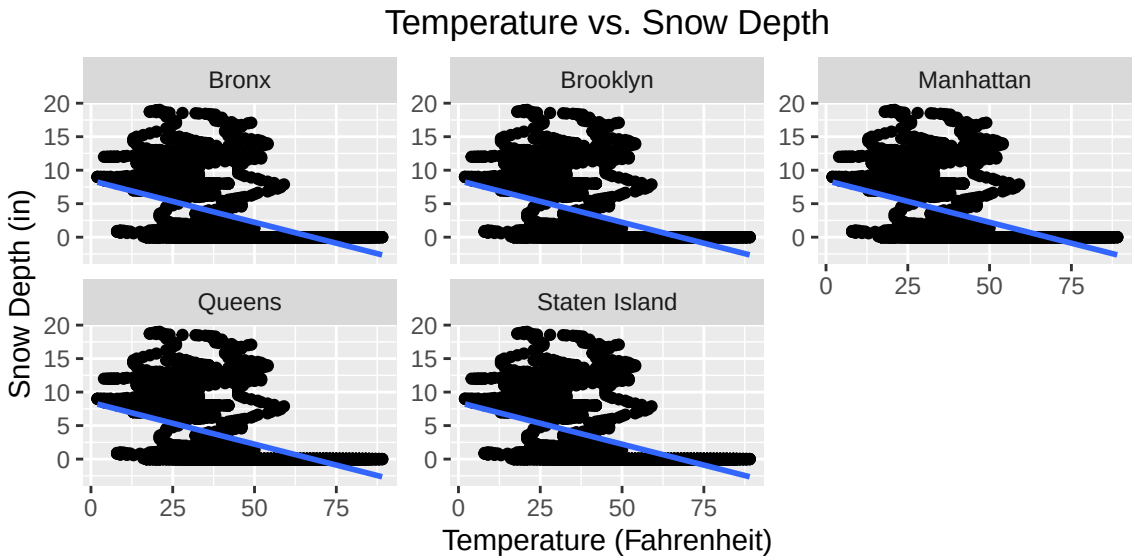

Temperature vs. Pickups by Borough
Bronx
Brooklyn
Manhattan
Queens
Staten Island
Pickups
Temperature (Fahrenheit)

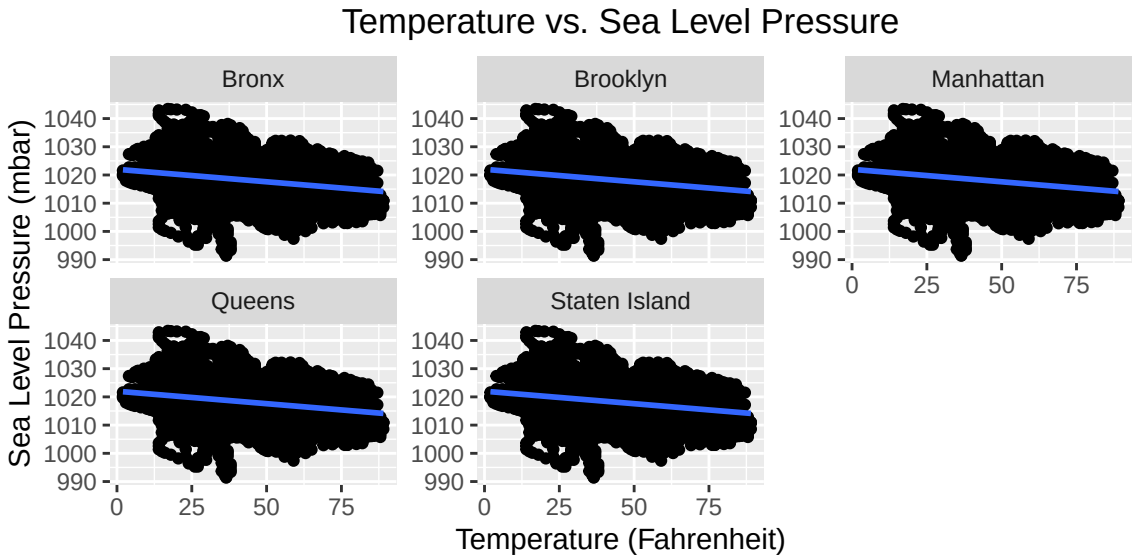

Snow Depth vs. Pickups by Borough
Bronx
Brooklyn
Manhattan
Queens
Staten Island
Pickups
Snow Depth (in)

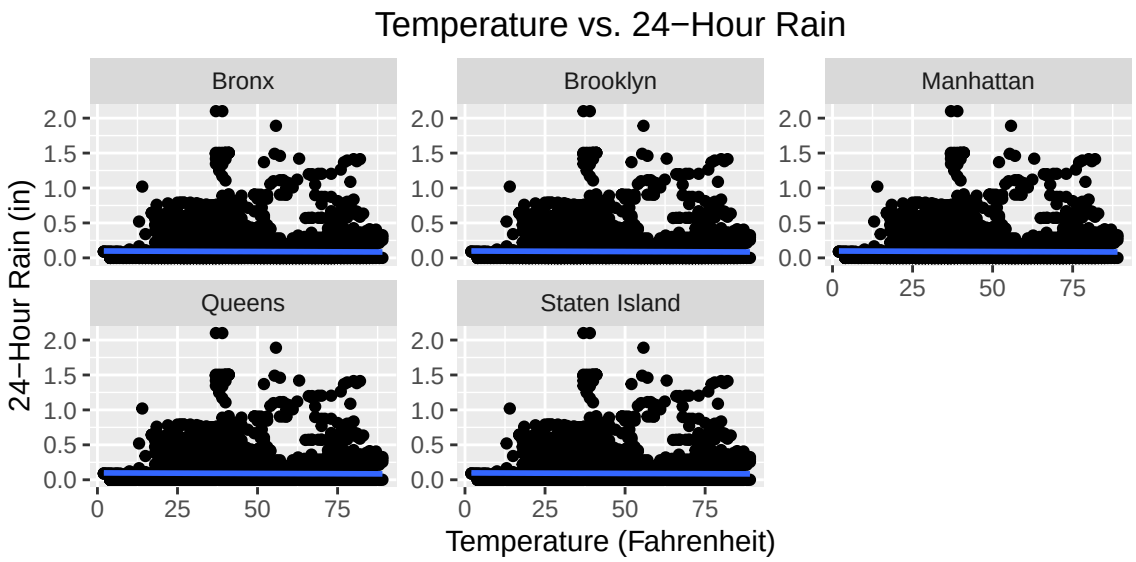

Sea Level Pressure vs. Pickups by Borough
Bronx
Brooklyn
Manhattan
Queens
Staten Island
Pickups
Sea Level Pressure (mbar)

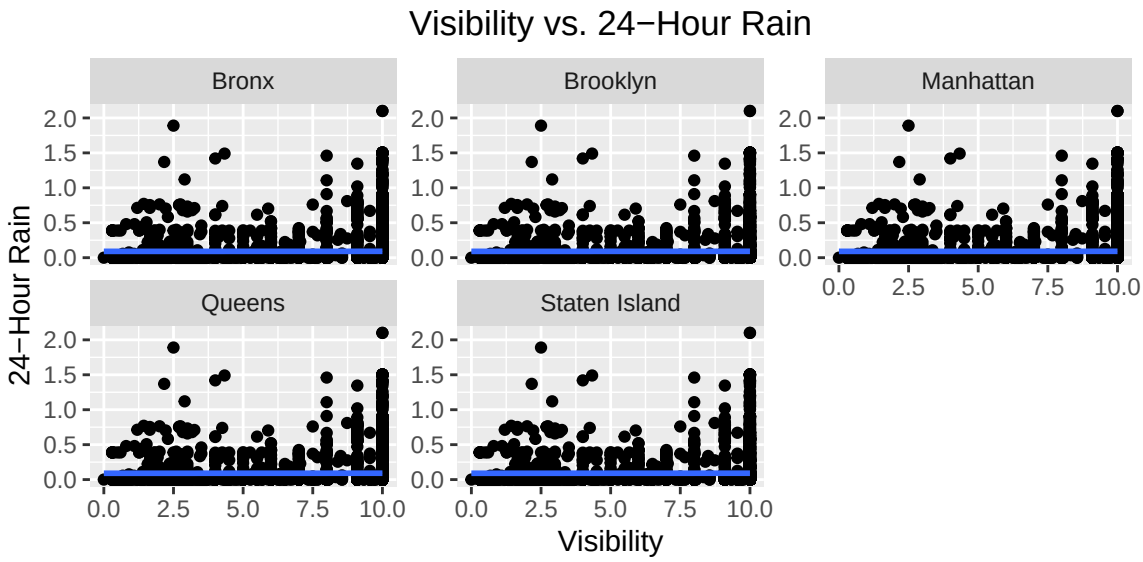


Appendix 1.4: Interaction Plots

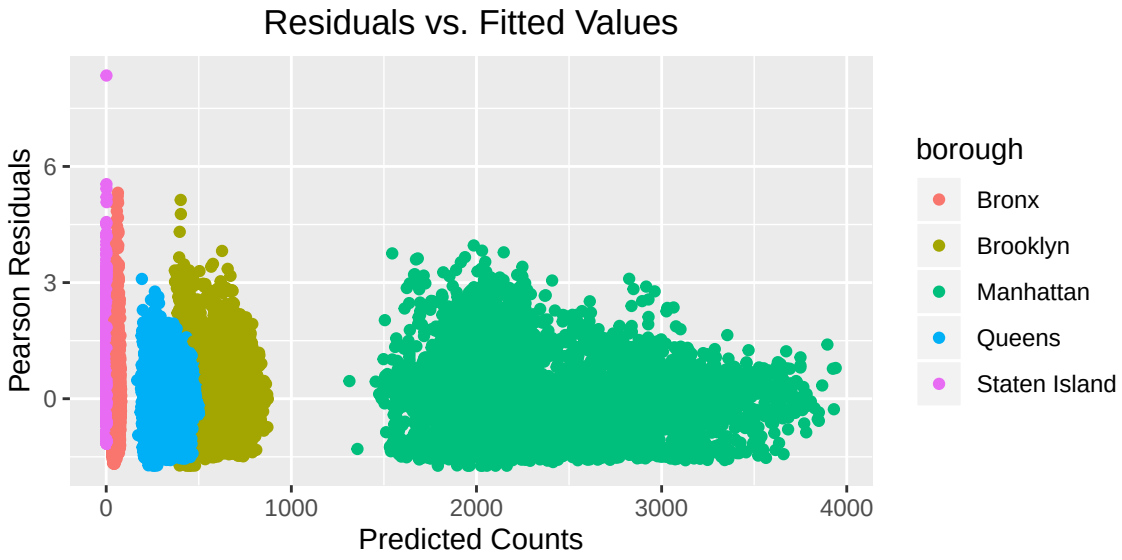


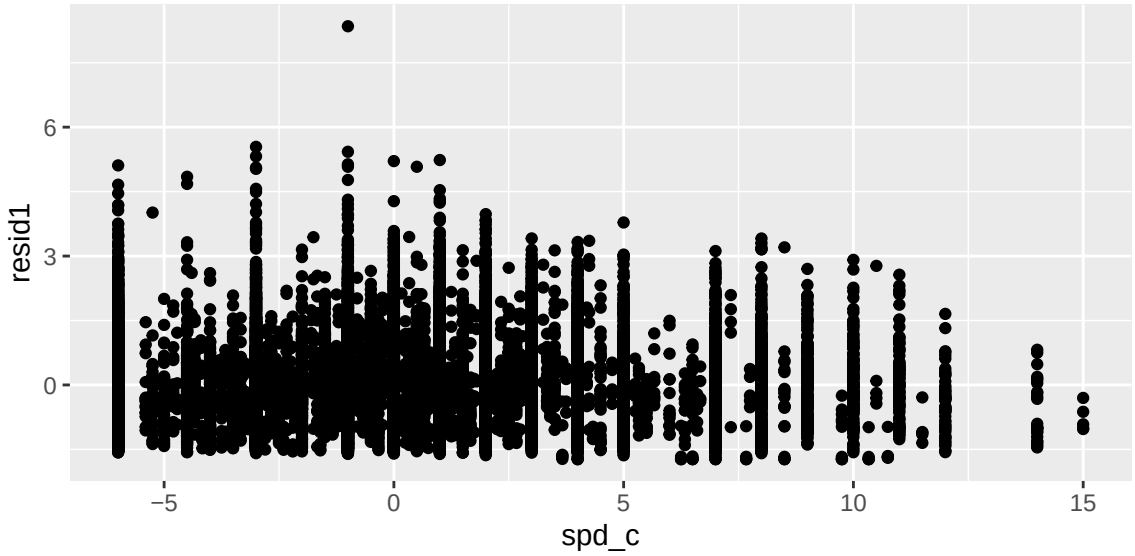

Temperature vs. Sea Level Pressure
Bronx
Brooklyn
Manhattan
Queens
Staten Island
Sea Level Pressure (mbar)
1040
1030
1020
1010
1000
990
0
25
50
75
Temperature (Fahrenheit)

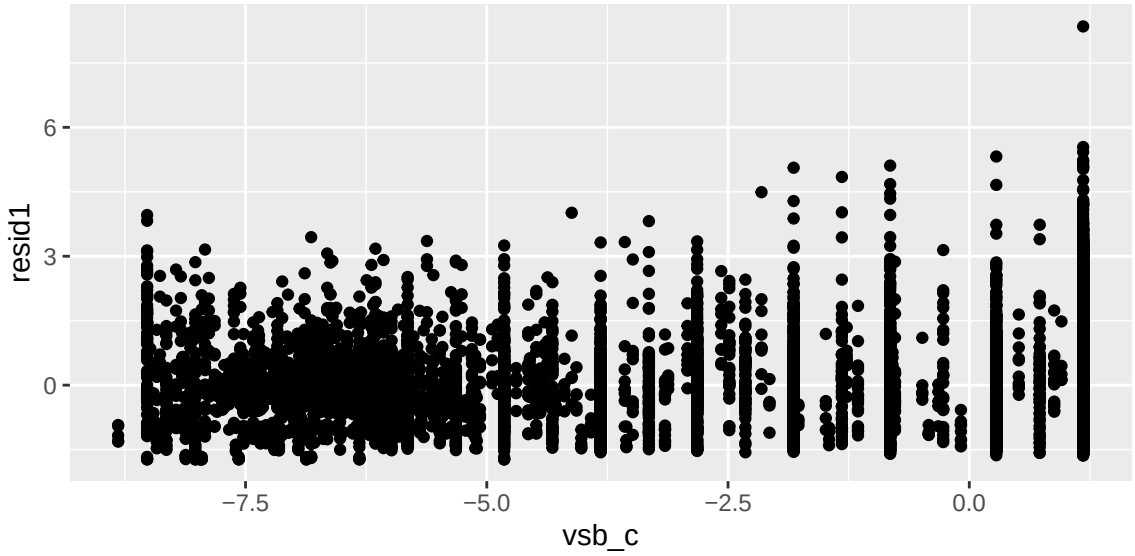

Temperature vs. 24−Hour Rain
Bronx
Brooklyn
Manhattan
Queens
Staten Island
24−Hour Rain (in)
2.0
1.5
1.0
0.5
0.0
0
25
50
75
Temperature (Fahrenheit)

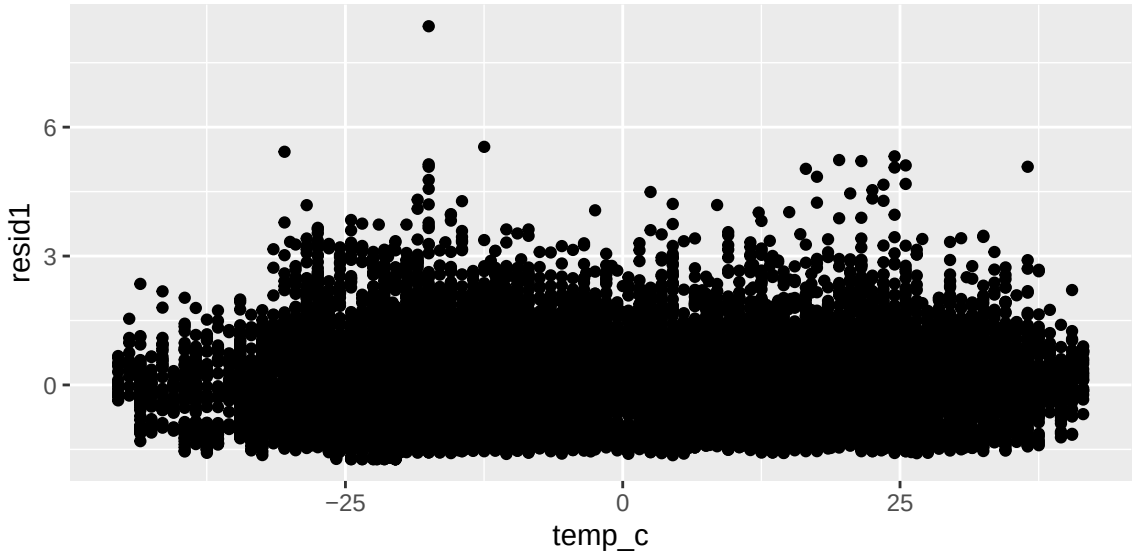


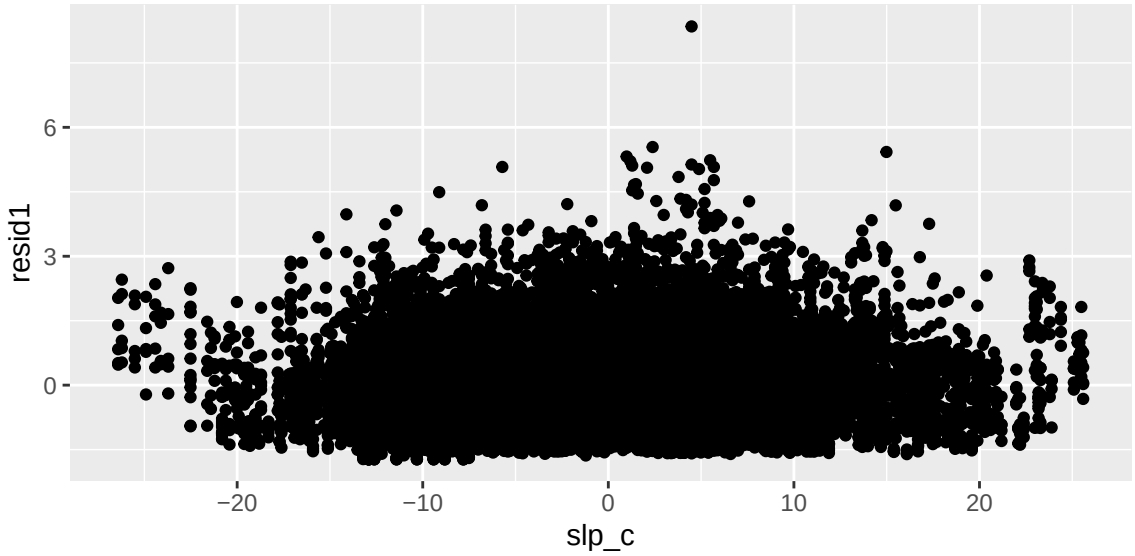
Appendix 2: Model Validation

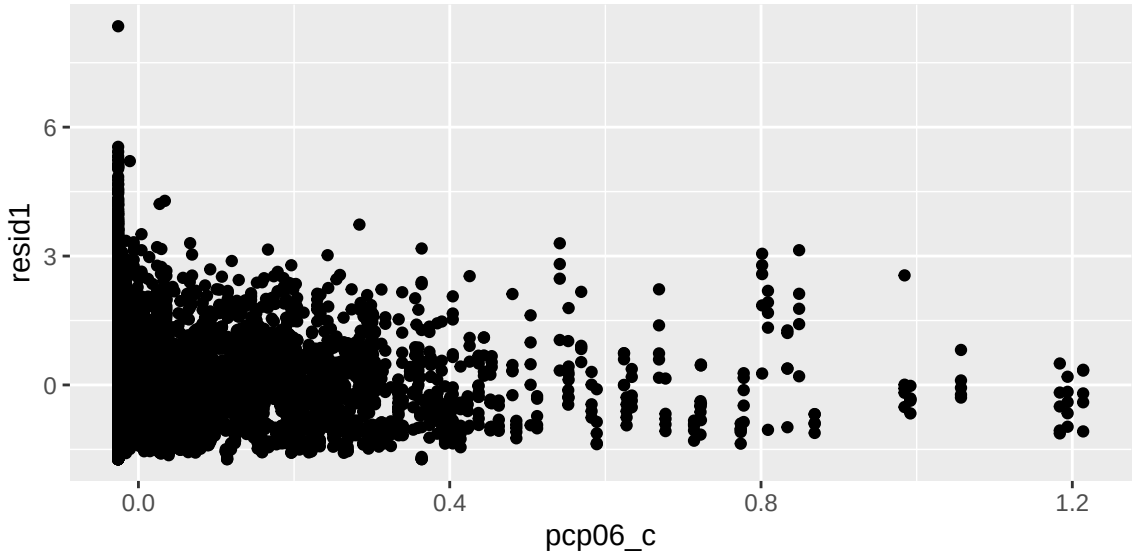


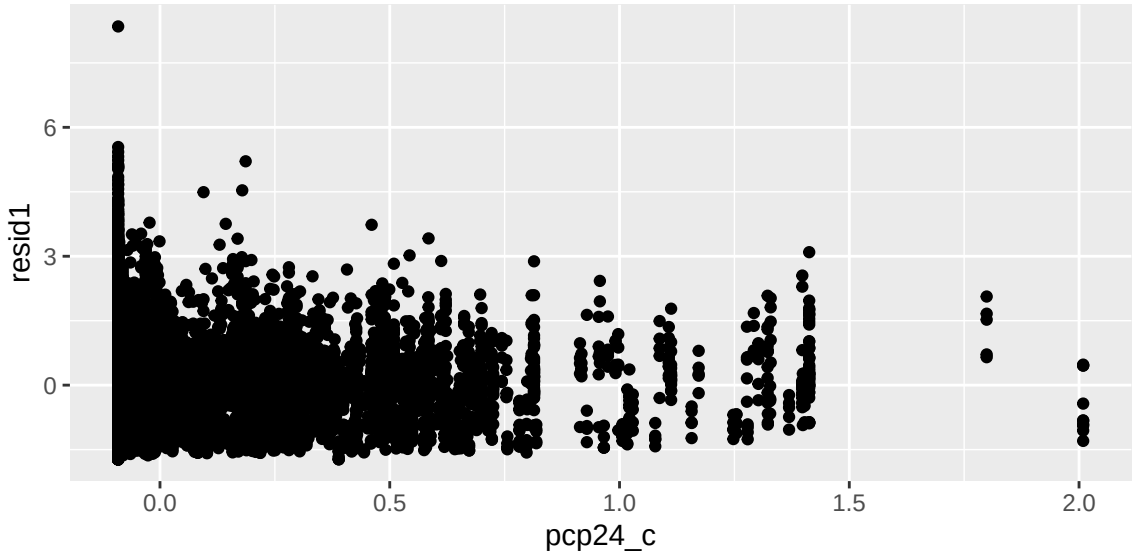

resid1
6
3
0
−5
0
5
10
15
spd_c

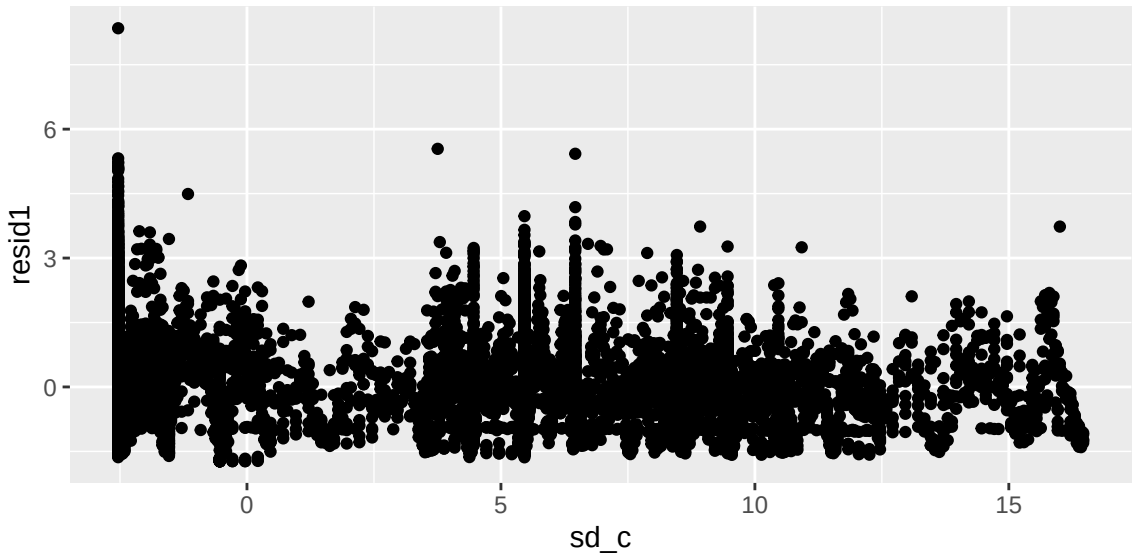

resid1
6
3
0
−7.5
−5.0
−2.5
0.0
vsb_c


resid1
6
3
0
−25
0
25
temp_c


resid1
6
3
0
−20
−10
0
10
20
slp_c


resid1
6
3
0
0.0
0.4
0.8
1.2
pcp06_c


resid1
6
3
0
0.0
0.5
1.0
1.5
2.0
pcp24_c


resid1
6
3
0
0
5
10
15
sd_c

Appendix 3: Code

```
knitr::opts_chunk$set(echo=FALSE, warning=FALSE, message=FALSE, fig.align = 'center',
                      fig.height = 3, fig.width = 6, fig.pos = "H")
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
library(gridExtra)
library(lattice)
library(xtable)
library(lme4)
library(lmerTest)
library(AER)
library(glmmTMB)
library(dplyr)
library(car)
library(MASS)
library(ggcorrplot)
library(xtable)
# Set directory
setwd("C:/Users/Varun/Documents/MIDS/Fall 2019/IDS 702 - Modeling and Representation of Data/Projects/F

# Import data
uber_og <- read.csv("uber_nyc_enriched.csv", header = T, sep = ",")

# Look at rows with NA as borough
uber_na <- uber_og[is.na(uber_og['borough']),]
summary(uber_na) # Max counts is 11 so remove

# Filtered dataframe without the NAs for borough
uber <- na.omit(uber_og)

# Remove EWR (max counts is 2)
uber <- uber[uber['borough'] != 'EWR',]

# Structure of data
str(uber)
summary(uber)

# Mean center variables on copy of original dataframe
uber2 <- uber
uber2$spd_c <- uber2$spd - mean(uber2$spd)
uber2$vsb_c <- uber2$vsb - mean(uber2$vsb)
uber2$temp_c <- uber2$temp - mean(uber2$temp)
uber2$dewp_c <- uber2$dewp - mean(uber2$dewp)
uber2$slp_c <- uber2$slp - mean(uber2$slp)
uber2$pcp01_c <- uber2$pcp01 - mean(uber2$pcp01)
uber2$pcp06_c <- uber2$pcp06 - mean(uber$pcp06)
uber2$pcp24_c <- uber2$pcp24 - mean(uber2$pcp24)
uber2$sd_c <- uber2$sd - mean(uber2$sd)

# # Standardize continuous variables for scaling
# uber_new <- uber
# uber_new$spd_c <- scale(uber_new$spd)
```

```r
# uber_new$vsb_c <- scale(uber_new$vsb)
# uber_new$temp_c <- scale(uber_new$temp)
# uber_new$dewp_c <- scale(uber_new$dewp)
# uber_new$slp_c <- scale(uber_new$slp)
# uber_new$pcp01_c <- scale(uber_new$pcp01)
# uber_new$pcp06_c <- scale(uber_new$pcp06)
# uber_new$pcp24_c <- scale(uber_new$pcp24)
# uber_new$sd_c <- scale(uber_new$sd)
#
# # Correlation matrix
# cor1 <- cor(uber_new[,c(14:22)]) # Dew point and temperature are highly correlated: remove dew point
# ggcorrplot(cor1, hc.order = TRUE, type = "lower", lab = TRUE)
# cor2 <- cor(uber2[,c(14:22)])
# ggcorrplot(cor2, hc.order = TRUE, type = "lower", lab = TRUE)

# Correlation matrix
cor2 <- cor(uber2[,c(14:22)])
ggcorrplot(cor2, hc.order = TRUE, type = "lower", lab = TRUE)
ggplot(uber, aes(x = borough, y = pickups, fill = borough)) +
  geom_boxplot() +
  labs(title = "Distribution of Pickups per Borough", x = "Borough",
       y = "Pickups") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")
# Holiday
#table(uber$hday)
ggplot(uber, aes(x = hday, y = pickups, fill = borough)) +
  facet_wrap(~borough, scales = "free_y") +
  geom_boxplot() +
  labs(title = "Pickups during Holidays by Borough", x = "Holiday?",
       y = "Pickups") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")
# Precipitation 24-hour
rain_plot <- ggplot(uber, aes(x = pcp24, y = pickups)) +
  facet_wrap(~borough, scales = "free_y") +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "24-Hour Liquid Precipitation vs. Pickups by Borough", x = "Rain (in)",
       y = "Pickups") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")
rain_plot
# Holiday
#table(uber$hday)
ggplot(uber, aes(x = hday, y = pickups, fill = borough)) +
  facet_wrap(~borough, scales = "free_y") +
  geom_boxplot() +
  labs(title = "Pickups during Holidays by Borough", x = "Holiday?",
       y = "Pickups") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")

# Precipitation
ggplot(uber, aes(x = pcp01, y = pickups)) +
  facet_wrap(~borough, scales = "free_y") +
  geom_point() +
```

```
  geom_smooth(method = "lm") +
  labs(title = "1-Hour Liquid Precipitation vs. Pickups by Borough", x = "Snow Depth (in)",
       y = "Pickups") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")

ggplot(uber, aes(x = pcp06, y = pickups)) +
  facet_wrap(~borough, scales = "free_y") +
  geom_point() + geom_smooth(method = "lm") +
  labs(title = "6-Hour Liquid Precipitation vs. Pickups by Borough", x = "Snow Depth (in)",
       y = "Pickups") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")

ggplot(uber, aes(x = pcp24, y = pickups)) +
  facet_wrap(~borough, scales = "free_y") +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "24-Hour Liquid Precipitation vs. Pickups by Borough", x = "Snow Depth (in)",
       y = "Pickups") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")

# Visibility
ggplot(uber, aes(x = vsb, y = pickups)) +
  facet_wrap(~borough, scales = "free_y") +
  geom_point() + geom_smooth(method = "lm") +
  labs(title = "Visibility vs. Pickups by Borough", x = "Visibility (miles)",
       y = "Pickups") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")

# Temperature
ggplot(uber, aes(x = temp, y = pickups)) +
  facet_wrap(~borough, scales = "free_y") +
  geom_point() + geom_smooth(method = "lm") +
  labs(title = "Temperature vs. Pickups by Borough", x = "Temperature (Fahrenheit)",
       y = "Pickups") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")

# Snow depth (inches)
snow_plot <- ggplot(uber, aes(x = sd, y = pickups)) +
  facet_wrap(~borough, scales = "free_y") +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Snow Depth vs. Pickups by Borough", x = "Snow Depth (in)",
       y = "Pickups") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")
snow_plot

# Sea level pressure
ggplot(uber, aes(x = slp, y = pickups)) +
  facet_wrap(~borough, scales = "free_y") +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Sea Level Pressure vs. Pickups by Borough", x = "Sea Level Pressure (mbar)",
       y = "Pickups") +
```

```
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")

# Wind speed
ggplot(uber, aes(x = spd, y = pickups)) +
  facet_wrap(~borough, scales = "free_y") +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Wind speed vs. Pickups by Borough", x = "Wind Speed (mph)",
       y = "Pickups") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")
# Snow depth and temp
ggplot(uber, aes(x = temp, y = sd)) +
  facet_wrap(~borough, scales = "free_y") +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Temperature vs. Snow Depth", x = "Temperature (Fahrenheit)",
       y = "Snow Depth (in)") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")

# Sea level pressure and temperature
ggplot(uber, aes(x = temp, y = slp)) +
  facet_wrap(~borough, scales = "free_y") +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Temperature vs. Sea Level Pressure", x = "Temperature (Fahrenheit)",
       y = "Sea Level Pressure (mbar)") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")

# Temperature and precipitation
ggplot(uber, aes(x = temp, y = pcp24)) +
  facet_wrap(~borough, scales = "free_y") +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Temperature vs. 24-Hour Rain", x = "Temperature (Fahrenheit)",
       y = "24-Hour Rain (in)") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")

# Visibility and precipitation
ggplot(uber, aes(x = vsb, y = pcp24)) +
  geom_point() +
  facet_wrap(~borough, scales = "free_y") +
  geom_smooth(method = "lm") +
  labs(title = "Visibility vs. 24-Hour Rain", x = "Visibility",
       y = "24-Hour Rain") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")
# Dipsersion check
dispersionstats <- uber %>% group_by(borough) %>% summarise(means = mean(pickups),variances = var(picku
dispersionstats # Overdispersed

# # Poisson models
# model1 <- glm(pickups~hday+borough, data = uber, family = poisson)
# summary(model1)
#
```

```
# model2 <- glm(pickups~hday+borough+spd_c+vsb_c+temp_c+slp_c+pcp01_c+pcp06_c+pcp24_c+sd_c,
#               data = uber2, family=poisson)
#
# dispersiontest(model2, trafo = 1)
# summary(model2)
#
# # Quasipoisson - scaled
# model3 <- glm(pickups~hday+borough+spd_c+vsb_c+temp_c+dewp_c+slp_c+pcp01_c+pcp06_c+pcp24_c+sd_c,
#               data = uber_new, family=quasipoisson)
# summary(model3)
#
# model4 <- glm(pickups~hday+borough+spd_c+vsb_c+temp_c+slp_c+pcp01_c+pcp06_c+pcp24_c+sd_c,
#               data = uber_new, family=quasipoisson) # Remove dew point
# summary(model4)
#
# model4a <- glm(pickups~hday+borough+spd+vsb+temp+slp+pcp01+pcp06+pcp24+sd,
#               data = uber_new, family=quasipoisson) # Remove dew point; use nonstandardized
# summary(model4a)
#
# model5 <- glm(pickups~hday+borough+spd_c+vsb_c+temp_c+slp_c+pcp01_c+pcp06_c+pcp24_c+sd_c
#               + temp_c:sd_c,
#               data = uber_new, family=quasipoisson) # Interactions
# summary(model5)
#
# # Negative binomial
# model6 <- glm.nb(pickups~hday+borough+spd+vsb+temp+slp+pcp01+pcp06+pcp24+sd, data = uber_new)
# summary(model6)
#
#
# Stepwise selection
# null_model <- glm.nb(pickups~1,data = uber2)
# full_model <- glm.nb(pickups~hday+borough+spd_c+vsb_c+temp_c+slp_c+pcp01_c+pcp06_c+pcp24_c+sd_c,
#                      data = uber2)
#
# model_backward <- step(full_model, direction = "backward", trace = 0)
# summary(model_backward)
#
# # Test interactions
# model7 <- model_backward
# model7a <- glm.nb(formula = pickups ~ borough + spd_c + vsb_c + temp_c + slp_c +
#                     pcp06_c + pcp24_c + sd_c + temp_c:sd_c, data = uber2) # significant
# summary(model7a) # Main model
#
# model7b <- glm.nb(formula = pickups ~ borough + spd_c + vsb_c + temp_c + slp_c +
#                     pcp06_c + pcp24_c + sd_c + temp_c:sd_c + temp_c:spd_c, data = uber2)
# summary(model7b)
# anova(model7a, model7b, test= "Chisq")
## Hierarchical Models
# Poisson
# p1 <- glmer(pickups~spd_c+vsb_c+temp_c+slp_c+pcp01_c+pcp06_c+pcp24_c+sd_c+hday+(1|borough), data = ub
# summary(p1)
#
# p2 <- glmer(pickups~hday+temp_c+sd_c+(1|borough), data = uber_new, family=poisson(link = "log"),
```

```
#             control=glmerControl(optimizer="bobyqa", optCtrl = list(maxfun = 100000)))
# summary(p2)
#
# # Negative binomial
# p3 <- glmer.nb(pickups~spd_c+vsb_c+temp_c+dewp_c+slp_c+pcp01_c+pcp06_c+pcp24_c+sd_c+hday+(1|borough),
# summary(p3)
#
# p4 <- glmer.nb(pickups~pcp06_c + (1|borough), data = uber_new, control=glmerControl(optimizer="bobyqa
# summary(p4)
#
# p4a <- glmer.nb(pickups~pcp06_c + (1|borough), data = uber2, control=glmerControl(optimizer="bobyqa",
# summary(p4a)
#
# p5 <- glmer.nb(pickups~spd_c+vsb_c+temp_c+dewp_c+slp_c+pcp01_c+pcp06_c+pcp24_c+sd_c+hday+(1|borough),
# summary(p5) # Add control
#
# p6 <- glmer.nb(pickups~spd_c+vsb_c+temp_c+slp_c+pcp01_c+pcp06_c+pcp24_c+sd_c+hday+(1|borough), data =
# summary(p6) # Remove dew point
#
# p7 <- glmer.nb(pickups~spd_c+(1|borough), data = uber2, control=glmerControl(optimizer="bobyqa", optC
# summary(p7)

# # Using glmmTMB package
# p8 <- glmmTMB(pickups~hday + spd_c + vsb_c + temp_c + slp_c + pcp06_c + pcp24_c + sd_c + (1|borough),
# summary(p8)

#### FINAL MODEL - commented out to save on knitting time
#p9 <- glmmTMB(pickups~spd_c + vsb_c + temp_c + slp_c + pcp06_c + pcp24_c + sd_c +
#                + temp_c:sd_c + (1|borough), data = uber2, family = nbinom2)
#saveRDS(p9,"final")
#sum <- summary(p9)
readRDS("final")
#saveRDS(sum,"Final_summary")
readRDS("final_summary")
#

# Dotplot
lme4:::dotplot.ranef.mer(ranef(readRDS("final"))$cond)
# Fitted values vs. residuals
# NB Multilevel (p9)
resid1 <- resid(readRDS("final"), type = "pearson")
pred1 <- predict(readRDS("final"),type="response")
#
# NB Borough as Factor
# resid2 <- resid(model7a, type = "pearson")
# pred2 <- predict(model7a, type = "response")
#
#qplot(y=resid1, x=pred1,data=uber_new,col=borough, geom="point",
#       xlab = "Predicted Counts", ylab = "Pearson Residuals")
#
 qplot(y=resid1, x=pred1, data=uber2, col=borough, geom="point",
       xlab = "Predicted Counts", ylab = "Pearson Residuals",
       main = "Residuals vs. Fitted Values") + theme(plot.title = element_text(hjust = 0.5))
```

```r
# Poisson for test
# resid3 <- resid(model2, type = "pearson")
# pred3 <- predict(model2, type = "response")
# qplot(y=resid3, x=pred3,data=uber_new,col=borough, geom="point",
#       xlab = "Predicted Counts", ylab = "Pearson Residuals")

# Continuous vs. residuals
ggplot(uber2, aes(spd_c,resid1)) + geom_point() # Wind speed
ggplot(uber2, aes(vsb_c,resid1)) + geom_point() # Visibility
ggplot(uber2, aes(temp_c,resid1)) + geom_point() # Temperature
ggplot(uber2, aes(slp_c,resid1)) + geom_point() # Sea level pressure
ggplot(uber2, aes(pcp06_c,resid1)) + geom_point() # 6-hour rain
ggplot(uber2, aes(pcp24_c,resid1)) + geom_point() # 24-hour rain
ggplot(uber2, aes(sd_c,resid1)) + geom_point() # Snow depth
```