How Weather Conditions and Day of the Week Influence the Number of Bikes Crossing Manhattan Bridge?

Kyle Ke

11/27/2021

Abstract

The analysis aims to describe how weather conditions and day of the week impact the number of bicycles crossing Manhattan Bridge in NYC on a given day using Poisson regression models. The analysis concludes that precipitation and high temperature are the two biggest factors impacting the number of bikes crossing Manhattan bridge on a given day: As precipitation goes up, the number of bikes goes down, and as high temperature increases, the number of bikes also increases. The most significant interaction observed is between day of the week and high temperature.

Contents

Introduction

Every day, thousands of bikers bike across East River via the Manhattan Bridge in New York City. The East River Bicycle Crossing data set from (Kaggle (n.d.)) contains the number of bicycles crossing Manhattan Bridge each day for a stretch of 9 months. Factors such as high temperature, low temperature, date, and precipitation are included in the data set. This analysis investigates how weather conditions and day of the week influence the number of bikes crossing Manhattan Bridge each day using Poisson regression models. The significance level was adjusted to be 0.002 (0.05/28) after Bonferoni correction.

Methdology

Data Preprocessing

First, a factor representing day of week was extracted from the date column. Next,the continuous explanatory variables high temperature and low temperature were both transformed into categorical variables with three levels: *Low*, *Mid*, and *High*. The explanatory variable precipitation was then reduced from nine levels to four levels: *High Snow*, *Mid High*, *Mid Low*, and *Low*. The levels were created to ensure equal interval between levels. Please refer to the code in the appendix for detailed splitting of the factor levels.

Modeling

Poisson regression (i.e. generalized linear model with poisson link function) was chosen to model the relationship between bicycle counts (dependent variable), and day of the week, high temperature, low temperature, and precipitation (explanatory variables). Poisson regression has the following four assumptions (Legler and Roback (2021)):

1. Poisson Response: The response is a count per unit of time, described by a poisson distribution.
2. Independence: The observations must be independent of one another.
3. Mean = variance: By definition, the mean of a Poisson random variable must be equal to its variance.
4. Linearity: The log of mean rate must be a linear function of x.

The model building process starts with the full model (i.e. model with the highest number of interactions). The full model is then gradually reduced to the final model. The statistical significance of the explanatory variables are determined by comparing nested models using the Likelihood ratio test (also Chi-squared test in this case). The model reduction process follows the following principles:

- Only variables that are deemed significant by the Chi-squared test are kept.
- Only variables that explain a large amount of the variation in the total sum of squares are kept.
- Ensure the root-mean-squared residuals of the final model is acceptable compared to the root-mean-squared residuals of the full model.
- Variables with large deviance are moved to the beginning of the reduced model.
- Make sure the model has a decent R-squared value, but also avoid over-fitting.
- Co-linear variables are removed.
- Follow Occan's razor by keeping the model as simple as possible.

Lastly, the coefficients of the final model is transformed from log(mean rate) to percent change for interpretability.

EDA

Table 1: Factors and Levels

Factor_name	Factor_Levels
day_of_week	Sun Mon Tue Wed Thu Fri Sat
High_Temp	Low, Mid, High
Low_Temp	Low, Mid,
Precipitation	Dry-Trace, Mid Low, Mid High, High Snow

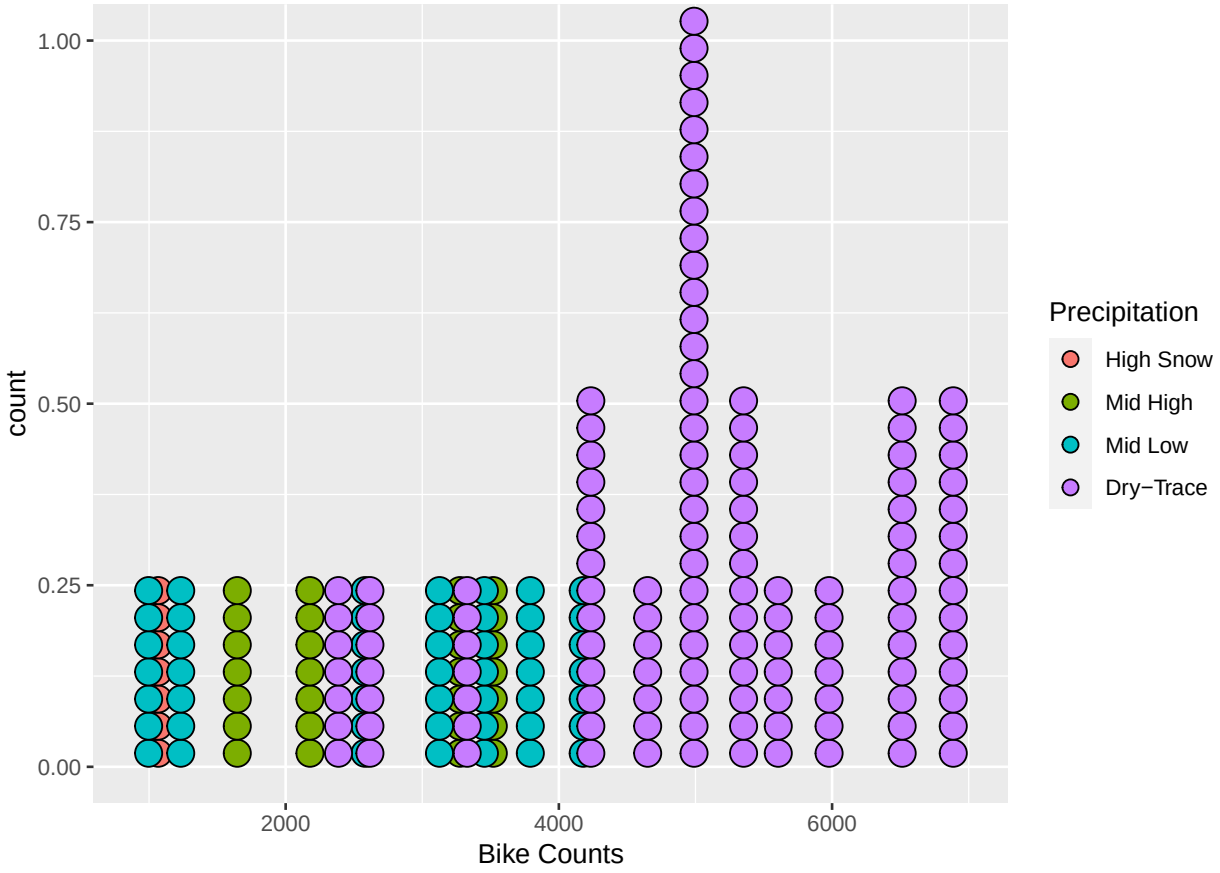

Figure 1: Bike Counts Dotplot by Precipitation

We observe from the dot plot of bike counts that when the bike counts is greater than approximately 4500, the precipitation level will always be dry to trace .

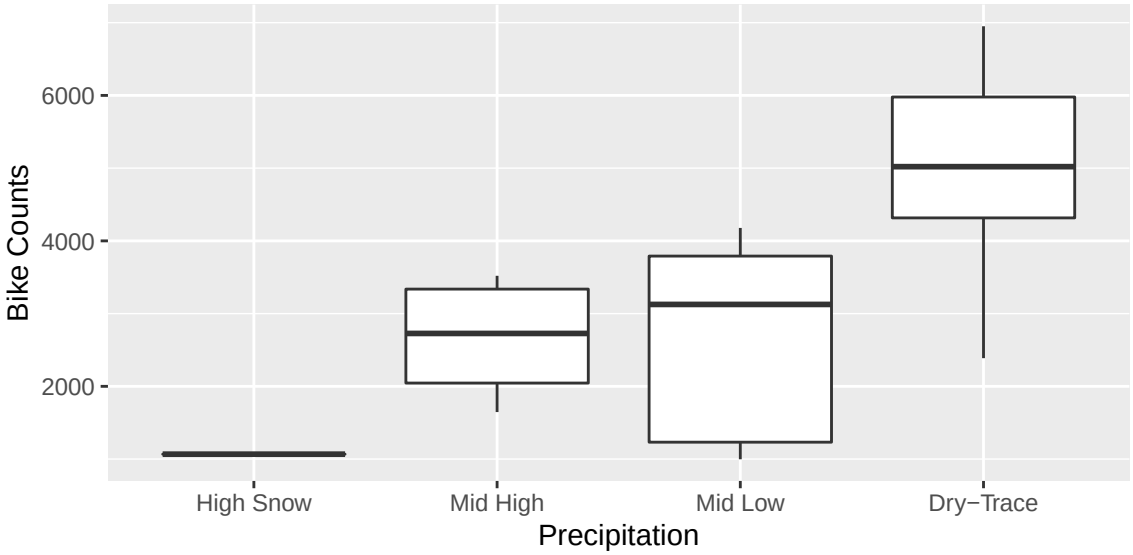

Figure 2: Bike Counts Boxplot by Precipitation

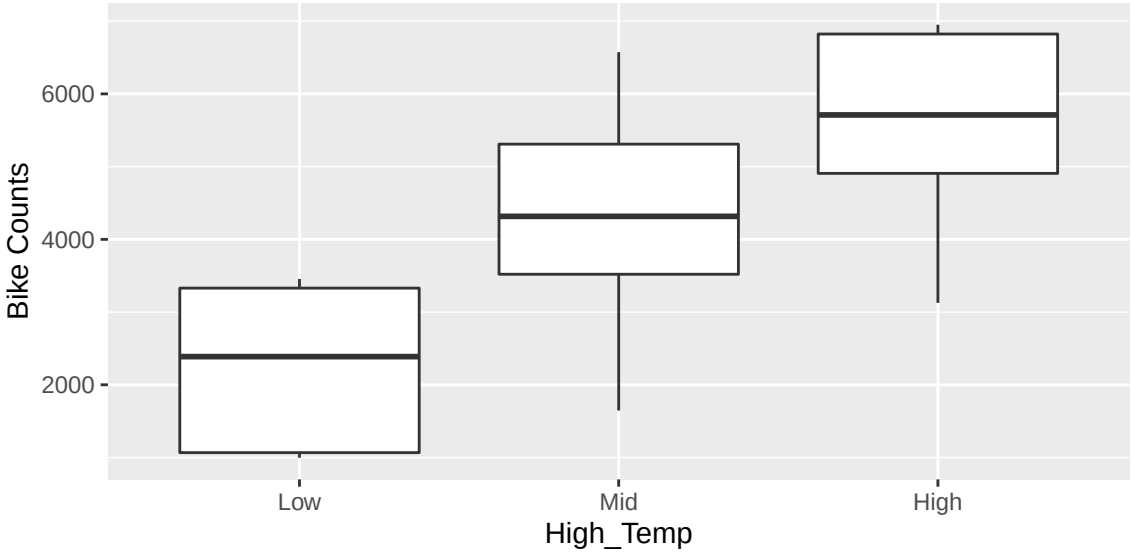

Figure 3: Bike Counts Boxplot by High Temp

Figure 2 shows that as precipitation decreases, bike counts increases, and figure 3 shows that as high temperature increases bike counts also increases.

Results

Full Model

Table 2: ANOVA of Full Model

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	209	164617.429	NA
day_of_week	6	26038.0036	203	138579.426	0
High_Temp	2	72112.5672	201	66466.859	0
Low_Temp	2	5883.7190	199	60583.140	0
Precipitation	3	41374.4840	196	19208.656	0
day_of_week:High_Temp	8	9933.6850	188	9274.971	0
day_of_week:Low_Temp	2	4125.2753	186	5149.695	0
day_of_week:Precipitation	1	595.4577	185	4554.238	0
High_Temp:Low_Temp	0	0.0000	185	4554.238	NA
High_Temp:Precipitation	1	483.0675	184	4071.170	0
Low_Temp:Precipitation	0	0.0000	184	4071.170	NA
day_of_week:High_Temp:Low_Temp	0	0.0000	184	4071.170	NA
day_of_week:High_Temp:Precipitation	0	0.0000	184	4071.170	NA
day_of_week:Low_Temp:Precipitation	0	0.0000	184	4071.170	NA
High_Temp:Low_Temp:Precipitation	0	0.0000	184	4071.170	NA
day_of_week:High_Temp:Low_Temp:Precipitation	0	0.0000	184	4071.170	NA

The ANOVA table of the full model shows the main effects: precipitation, day of week, high temperature, and low temperature are all statistically significant with p-value of 0. The interaction terms high temperature and day of week, low temperature and day of week, and precipitation and weekday are all statistically significant with p-value of 0. Within the statistically significant factors, day of week explains 15.8% of the variation in the total sum of squares. High temperature, adjusted for day of the week, explains 43.8% of the variation in the total sum of squares. And precipitation, adjusted for day of the week, high temperature, and low temperature, explains 25.1% of the variation in the total sum of squares.

The R-squared value of the full model is calculated to be 0.972,meaning 97.2% of the variation in bike counts is explained by the full model. In addition, the root-mean-squared residuals is calculated to be 273.287, meaning on average, predicted daily bike counts from the full model is 273.287 counts away from the actual value of daily bike counts.

Following the model reduction principles described in the methods section, the final model only includes precipitation, high temperature, day of the week, and the interaction between high temperature and day of the week. Note that low temperature is dropped because it is co-linear with high temperature. Please see the appendix for the detailed model reduction process and the in-between models.

Final Model

The final model is split into two models in order to interpret both the main effects and the interaction effects: The first model includes only the main effects of precipitation, high temperature, and day of the week. The second model includes all of the aforementioned main effects and an added interaction term between high temperature and day of the week.

Table 3: ANOVA of Final Model (main effects only)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	209	164617.43	NA
Precipitation	3	87072.30	206	77545.13	0
High_Temp	2	41681.36	204	35863.77	0
day_of_week	6	11174.47	198	24689.30	0

The ANOVA table of the final model (main effects only) shows that precipitation alone account for 52.9% of the variation in the total sum of squares. High temperature, adjusted for precipitation, account for 25.3% of the variation in the total sum of squares.The R-squared value is calculated to be 0.85,meaning 85% of the variation in bike counts is explained by the main-effects-only final model. In addition, the root-mean-squared residuals is calculated to be 618.636, meaning on average, predicted daily bike counts from the main-effects-only final model is 618.636 counts away from the actual value of daily bike counts. The difference between root-mean-squared residuals of the full model and the main-effects-only final model is around 345 bikes.

Table 4: Final Model (main effects only) Coeffcients as Percent Change

	Percent_Change
(Intercept)	106600.00
PrecipitationMid High	69.37
PrecipitationMid Low	102.69
PrecipitationDry-Trace	219.17
High_TempMid	77.06
High_TempHigh	97.71
day_of_weekTue	-7.69
day_of_weekWed	-2.75
day_of_weekThu	-18.25
day_of_weekFri	-9.95
day_of_weekSat	-26.35
day_of_weekSun	-27.78

From the coefficient table of the main effects only final model, we see that comparing to the high-snow level of precipitation, the average bike counts increases by 69.37% when the precipitation level is mid-high, increases by 102.69% when the precipitation level is mid-low, and increases by 219.17% when the precipitation level is dry-trace, holding all else constant. In addition, we observe that comparing to the low level of high temperature, the average bike counts increase by 77.06% when the high temperature level is mid and increases by 97.71% when the high temperature level is high, holding all else constant. All of the days of the week seem to decrease the average bike counts compared to Monday.

Table 5: ANOVA of Final Model (with interaction)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	209	164617.429	NA
Precipitation	3	87072.30	206	77545.129	0
High_Temp	2	41681.36	204	35863.771	0
day_of_week	6	11174.47	198	24689.305	0
High_Temp:day_of_week	8	14723.88	190	9965.427	0

Adding the interaction effect between day of the week and high temperature, the R-squared of the final model (with interaction) increased from 0.85 (main effects only) to 0.939. The root-mean-squared residuals is now 398.784, which only differs from the root-mean-squared residuals of the full model by 125.497.

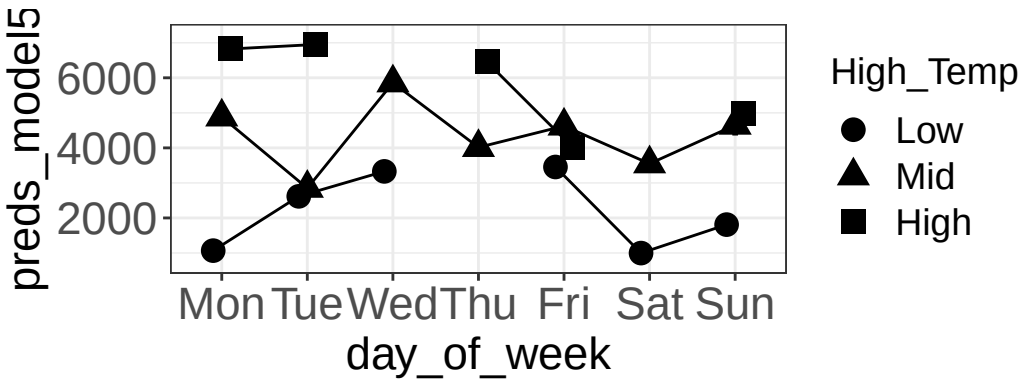

Figure 4: Day of the Week and High Temp Interaction 1

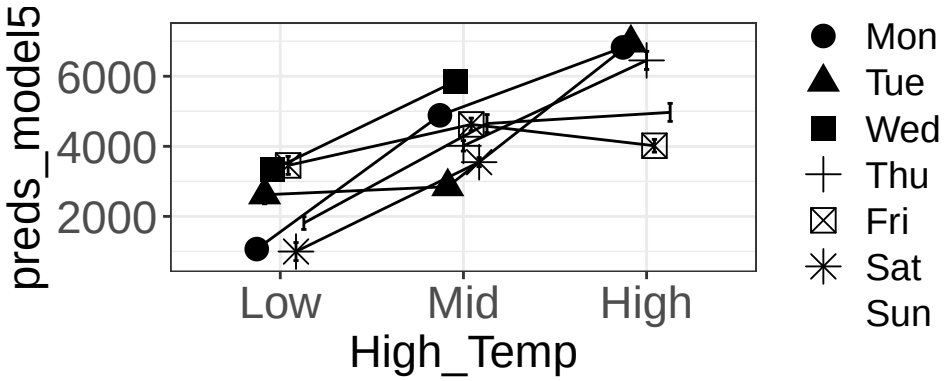

Figure 5: Day of the Week and High Temp Interaction 2

From the interaction plot 1 between day of the week and high temperature, we observe that when the high temperature is at the mid level, predicted bike counts tend to decrease drastically from Monday to Tuesday. However, when the high temperature is at the low level, predicted bike counts increases drastically from Monday to Tuesday. In addition, when the high temperature is at the high level, predicted bike counts decreases from Thursday to Friday. On the other hand, when the high temperature is at the mid level, predicted bike counts increases from Thursday to Friday.

From interaction plot 2 between day of the week and high temperature, we observe that on Fridays, predicted bike counts tend to remain the same regardless of high temperature. On Tuesdays, predicted bike counts tends to only increase a little bit from the low level to mid level, but increases drastically from the mid level to high level. On Sundays, predicted bike counts increases greatly from low level to mid level of high temperature, but remains relatively the same from mid level to high level.

Checking Assumptions

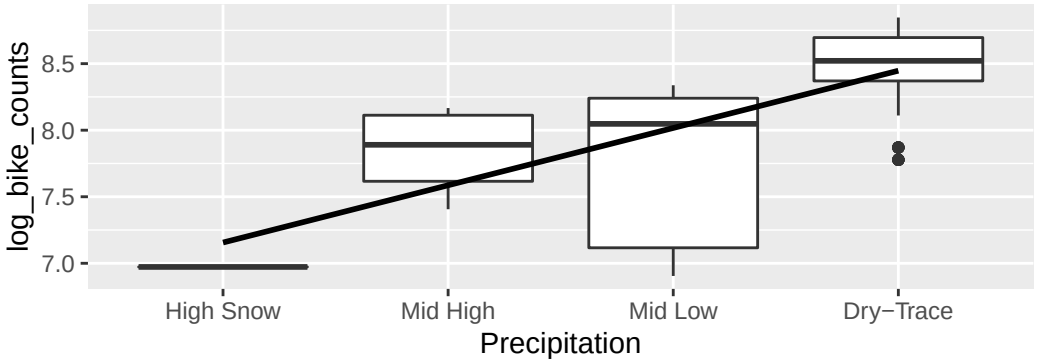

Figure 6: Log(Bike Counts) Boxplot by Precipitation

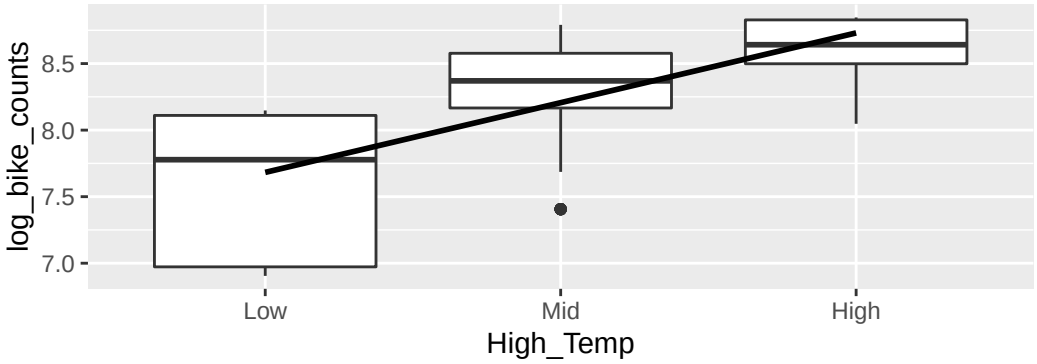

Figure 7: Log(Bike Counts) Boxplot by High Temp

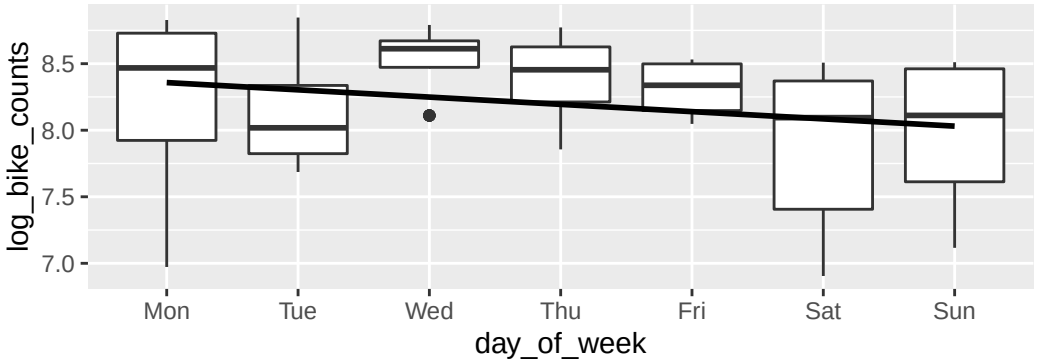

Figure 8: Log(Bike Counts) Boxplot by Day of the Week

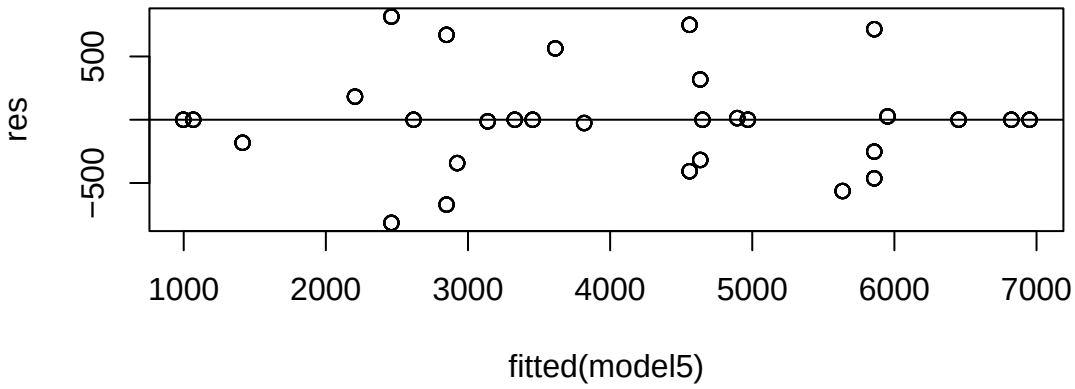

Figure 9: Resid vs Fitted (Final Model with Interaction)

1. Poisson Response: Bike counts is count per day.
2. Independence: NOT MET. From the residual vs fitted value plot, we do not see any violation of independence. However, data is collected on consecutive dates. Thus, there has to be some kind of time correlation between data points.
3. Mean = variance: NOT MET. Mean 4050 does not equal variance 2.906109×10^6
4. Linearity: From the box plots between log(bike counts) and the variables we see the linearity assumption is met.

Conclusions

Precipitation is the biggest factor that influences the number of bikes crossing Manhattan Bridge. Generally speaking, as precipitation goes down, the number of bikes increases. The second biggest factor is high temperature. Generally speaking, as high temperature increases the number of bikes also increases. Together, these two factors explain over two-thirds of the variation in the total sum of squares. Moreover, bike counts start out high on Monday, and drop drastically on Tuesday, climb back up on Wednesday, and then decline towards the weekends. The greatest interaction observed is between high temperature and day of the week. This interaction effect is explained in the results section. Thus, the analysis achieved its objective of explaining how weather conditions and day of the week influence the number of bikes crossing Manhattan Bridge in a given day. Although we did not test our final model on a test set, we have reasons to believe the final model (with interaction) will predict well on unseen data because it has a high R-squared of 0.939.

Recommedations for Future Studies

The data points are gathered on consecutive dates. Further studies could investigate the time series aspect of the data. This is a hole in the current analysis that needs to be filled. Furthermore, the data only spans a period of 9 months. It'd be more ideal if a full year of data can is gathered.

References

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Helwig, Nathaniel. 2017. "Permutation Tests." University of Minnesota Twin Cities. https://statmoodle.byu.edu/pluginfile.php/168/mod_resource/content/2/perm-Notes.pdf.

Kaggle. n.d. "New York City - East River Bicycle Crossings." *NYC Dept of Transportation*. kaggle. https://www.kaggle.com/new-york-city/nyc-east-river-bicycle-crossings.

Legler, Julie, and Paul Roback. 2021. "Beyond Multiple Linear Regression." *Beyond Multiple Linear Regression*. CRC Press. https://bookdown.org/roback/bookdown-BeyondMLR/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2021. *Rvest: Easily Harvest (Scrape) Web Pages*. https://CRAN.R-project.org/package=rvest.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2021. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Jim Hester. 2021. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Wuertz, Diethelm, Tobias Setz, and Yohan Chalabi. 2020. *fBasics: Rmetrics - Markets and Basic Statistics*. https://www.rmetrics.org.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/%0A%20%20%20%209781466561595.

———. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

———. 2021a. *formatR: Format r Code Automatically*. https://github.com/yihui/formatR.

———. 2021b. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Appendix

Code

```
knitr::opts_chunk$set(message = FALSE)
library(dplyr)
library(ggplot2)
library(chron)
library(fBasics)
library(knitr)
library(formatR)
library(broom)
library(jtools)
library(readr)
library(kableExtra)
library(data.table)
library(emmeans)

source('gg_interaction_function.R')
knitr::write_bib(c('rvest','dplyr','ggplot2','fBasics','knitr','formatR','janitor','readr'),
                 file='packages.bib',width = 60)
df <-readr:: read_csv("nyc-east-river-bicycle-counts.csv")
#creating a new variable: day of week
df$Date <- as.Date(df$Date)
df$day_of_week <- weekdays(df$Date,abbreviate=TRUE)
df$day_of_week <- factor(df$day_of_week,levels= c( "Mon",
    "Tue", "Wed", "Thu", "Fri", "Sat","Sun"))
#We only care about
df <- df[,c(4,5,6,8,12)]
colnames(df)[4] <- "Bike Counts"


df <- df %>% mutate(Precipitation=case_when(
    Precipitation %in% c("0","T") ~ 'Dry-Trace',
    Precipitation %in%  c("0.01","0.05","0.09") ~ 'Mid Low',
    Precipitation %in% c("0.15","0.16","0.2","0.24") ~ 'Mid High',
    Precipitation %in% c("0.47 (S)") ~ "High Snow"

))


df <- df %>% mutate(Low_Temp=case_when(
     `Low Temp (°F)`>= 26.1 & `Low Temp (°F)` <= 39.9 ~ 'Low',
        `Low Temp (°F)`> 39.9 & `Low Temp (°F)` <= 52.0 ~ 'Mid',
    `Low Temp (°F)`> 52.0 & `Low Temp (°F)` <= 66.0 ~ 'High',
 ))

df <- df %>% mutate(High_Temp=case_when(
     `High Temp (°F)`>= 39.9 & `High Temp (°F)` < 55.0 ~ 'Low',
        `High Temp (°F)`>= 55.0 & `High Temp (°F)` < 71.1 ~ 'Mid',
     `High Temp (°F)`>= 71.1 & `High Temp (°F)` <= 81.0 ~ 'High',
 ))
df$Precipitation = factor(df$Precipitation,levels=c("High Snow","Mid High","Mid Low","Dry-Trace"))
df$Precipitation=relevel(df$Precipitation,ref="High Snow")
```

```r
df$Low_Temp= factor(df$Low_Temp,levels=c("Low","Mid","High"))
df$High_Temp=factor(df$High_Temp,levels=c("Low","Mid","High"))
df$High_Temp=relevel(df$High_Temp,ref="Low")
df$Low_Temp=relevel(df$Low_Temp,ref="Low")
df$day_of_week=factor(df$day_of_week)
df$day_of_week=relevel(df$day_of_week,ref="Mon")
sum(is.na.data.frame(df))
Factor_name = c("day_of_week","High_Temp","Low_Temp","Precipitation")
Factor_Levels = c("Sun Mon Tue Wed Thu Fri Sat","Low, Mid, High","Low, Mid, ","Dry-Trace, Mid Low, Mid H
factor_and_levels= data.frame(Factor_name,Factor_Levels)
knitr::kable(factor_and_levels,caption = "Factors and Levels",) %>% kable_styling(position = "center",la
ggplot(df, aes(x = `Bike Counts`, fill = Precipitation)) + geom_dotplot(binaxis = 'x', stackdir = 'up')
ggplot(df, aes(x = Precipitation, y = `Bike Counts`)) + geom_boxplot()

ggplot(df, aes(x = High_Temp, y = `Bike Counts`)) + geom_boxplot()

model1 <- glm(`Bike Counts` ~ (day_of_week+`High_Temp`+`Low_Temp`+Precipitation)^4,family="poisson", da
anova_model1 <- anova(model1,test='Chisq')
knitr::kable(anova_model1,caption = "ANOVA of Full Model") %>% kable_styling(position = "center",latex_
dayofweek_explains_1 = round(anova_model1$Deviance[2]/anova_model1$`Resid. Dev`[1],digits=3)
highTemp_explains_1 = round(anova_model1$Deviance[3]/anova_model1$`Resid. Dev`[1],digits=3)
precip_explains_1 = round(anova_model1$Deviance[5]/anova_model1$`Resid. Dev`[1],digits=3)
r_squared_1 = round(1-anova_model1$`Resid. Dev`[9]/anova_model1$`Resid. Dev`[1],digits = 3)
rmse_1 = round(sqrt(mean((residuals(model1,type='response')^2))),digits=3)
#plot(predict(model1,type='response'),predict(model1,type='response')-df$`Bike Counts` )
#Dropping all of the non-significant features and order the sequential model based on the features with

model2 <- glm(`Bike Counts` ~ `High_Temp`+Precipitation+day_of_week+day_of_week:Precipitation+day_of_wee
anova(model2,test='LRT')
#Dropping colinear terms (Low_Temp) with small deviance day_of_week:Low_Temp, Precipitation:day_of_week

model3 <- glm(`Bike Counts` ~ `High_Temp`+Precipitation+day_of_week+day_of_week:High_Temp,family="poisso
anova(model3,test='LRT')




#Switching the order of Precipitation and `High_Temp`

model4 <- glm(`Bike Counts` ~ Precipitation+`High_Temp`+day_of_week,family="poisson", data=df)
#plot(predict(model4,type='response'),predict(model4,type='response')-df$`Bike Counts` )
anova_model4 <- anova(model4,test='Chisq')
knitr::kable(anova_model4,caption = "ANOVA of Final Model (main effects only)") %>% kable_styling(positi
precipitation_explain_4 = round(anova_model4$Deviance[2]/anova_model4$`Resid. Dev`[1],digits=3)
high_temp_explain_4 = round(anova_model4$Deviance[3]/anova_model4$`Resid. Dev`[1],digits=3)
r_squared_4 = round(1-anova_model4$`Resid. Dev`[4]/anova_model4$`Resid. Dev`[1],digits = 3)
rmse_4 = round(sqrt(mean((residuals(model4,type='response')^2))),digits=3)
model4_coef = as.data.frame(round(100*(exp(coef(model4))-1),2))
colnames(model4_coef)[1] <- "Percent_Change"
knitr::kable(model4_coef,caption = "Final Model (main effects only) Coeffcients as Percent Change") %>%
model5 = glm(`Bike Counts` ~ Precipitation+`High_Temp`+day_of_week + `High_Temp`:day_of_week ,family="po
anova_model5 <- anova(model5,test='Chisq')
knitr::kable(anova_model5,caption = "ANOVA of Final Model (with interaction)")
```

```
#%>% kable_styling(position = "center",latex_options = "HOLD_position")
r_squared_5 = round(1-anova_model5$`Resid. Dev`[5]/anova_model4$`Resid. Dev`[1],digits = 3)
interaction_explains = round(anova_model5$Deviance[5]/anova_model5$`Resid. Dev`[1],digits=3)
rmse_5 = round(sqrt(mean((residuals(model5,type='response')^2))),digits=3)



preds_model5 <- predict(model5,type="response")
df_pred<- cbind(df, preds_model5 )
gg_interaction(x = c("day_of_week", "High_Temp"), y = "preds_model5", random = NULL, data = df_pred)



gg_interaction(x = c("High_Temp", "day_of_week"), y = "preds_model5", random = NULL, data = df_pred)

df$log_bike_counts = log(df$`Bike Counts`)
ggplot(df, aes(x = Precipitation, y = log_bike_counts)) + geom_boxplot()+geom_smooth(method = "lm", se=

ggplot(df, aes(x = High_Temp, y = log_bike_counts )) + geom_boxplot() + geom_smooth(method = "lm", se=F

ggplot(df, aes(x = day_of_week, y = log_bike_counts)) + geom_boxplot() + geom_smooth(method = "lm", se=

res <- resid(model5,type='response')
plot(fitted(model5), res)
abline(0,0)
Mean = round(mean(df$`Bike Counts`),digits=3)
variance = round(var(df$`Bike Counts`),digits=3)

knitr::opts_chunk$set(tidy.opts=list(width.cutoff=100), tidy=TRUE)
```