Two-Tired of COVID-19:
Impact of the COVID-19 Pandemic on Boston's Bikeshare Program

Christine Cai, Rucha Joshi, Vincent Pan, Grace Tian

---

1. Introduction

Background

An environmentally friendly mode of transportation, biking has become the quintessential 21st century sport that provides relief to our increasingly digital population to the breathtaking views of the outside world. With bike share hubs popping up all over the world, we set out to investigate bike share in our beloved city that we call home, Boston.

Bluebikes (originally known as Hubway) is a public bike share system in the Boston metropolitan area. It serves the municipalities of Boston, Brookline, Cambridge, Everett, and Somerville and offers its users a "fun and affordable way to get around" (Bluebikes) by letting them pick up and return a bike at any station for one-way rides, commuting, or leisure. Bluebikes has over 3,000 bikes across more than 300 stations in the 5 municipalities and as such is a very popular and growing bike share program.

Although Bluebikes has expanded its presence in the Boston community, the diffusion of COVID-19 has pivoted its functionality. Since the start of March 2020, many cities and states have implemented lockdown policies and stay-at-home orders, limiting outdoor activities and interactions. Many workplaces have closed their doors to their employees, asking them to work from home instead. Patrons of bars and restaurants have ceased visiting their favorite joints and where there was once noisy traffic is filled with a deafening silence. Once seen as an opportunity to go out and explore the outdoors, Bluebikes is now a wonderful excuse to simply get out of the house.

When a state of emergency was first declared, most people took lockdown procedures seriously due to the rise of infections and the uncertainty surrounding the risks of COVID-19. However, after some time people may have begun to doubt the necessity of social distancing and isolating. Adding to the loosening of early pandemic protocols this led to another wave of infections and another round of lockdowns. We wanted to examine the effect of the COVID-19 pandemic on a business that promotes individual outdoor activity (biking) and public transportation. First, we want to analyze the effects of COVID-19 on the frequency of Bluebikes rentals while controlling for other factors such as weather, which may impact the usage of Bluebikes. We then wanted to explore the idea of "lockdown fatigue" and returning to pre-lockdown rates of activity and usage even though regulations and limitations are still in effect. Will people be more inclined to "get out of the house" and bike after prolonged lockdown procedure? In the pandemic, are people still using Bikeshare in lieu of other public transportation for purposes like commuting, or are people largely biking for leisure?

Hypothesis

First, we hypothesize that the COVID-19 pandemic has negatively impacted usership of bikeshare. This may have consequently changed the primary users of Bluebikes. We plan to use predictive models to assess how bike usership has changed with LASSO, random forest, and mixed modeling techniques and see how they compare by considering MSEs and ANOVA testing.

Second, as lockdown has grown longer, the longing for life to go back to "normal" has increased. Our exploratory data analysis of bike share frequency over time brings up interesting questions about lockdown fatigue. Unlike biking in 2018 and 2019, the bike share frequency in 2020 has an unusual U shape. This begs the question - is bike share in 2020 better modeled by a quadratic U shape due to the lockdown orders? Or is this simply caused by weather effects that are similar across other years?

We hypothesize that there is lockdown fatigue, i.e. the longer the social distancing procedures are in place, the less people stay inside and bike more. Municipalities such as Cambridge or Brookline may also be different from that of Boston proper. Is there a difference in bike sharing between these districts? Let's find out.

2. Exploratory Data Analysis

Data Collection

To be able to compare bike sharing across summer months before and after the pandemic took place, we took a look at the months of March, April and May in 2018, 2019, and 2020 for a total of 9 months worth of data. We chose these months to hold as many factors constant as possible and to isolate the impact of COVID-19 -- the restrictions in Boston began in March 2020 and Phase 1 reopening was announced at the end of May, so Boston residents would largely have spent March through May of 2020 under lockdown procedures.

After downloading and combining the monthly datasets (50,000 - 200,000 entries each) we grouped by starting bike station and day to form an initial exploratory dataset *start3.t* (61,000 rows). Bluebikes also provides data detailing the district that a station is located. We created a categorical district variable to account for possible differences between each district. Two other groupings were then made: first, total rentals per district which we named *start4.t* (1465 rows) and second, total rentals per day, named *start2 (*275 rows). The main variable that we are choosing to predict will be the number of bike rentals, referred to throughout the rest of this paper as "frequency". In all three tables, totals were stored as a frequency variable, which will be the response variable of interest in our analysis. We also included a binary weekend predictor, as different demographics likely bike during weekdays rather than weekends: those who use Bluebikes to commute are more likely to bike during weekdays, whereas leisure bikers likely bike more over weekends.

Along with the Bluebikes rental data, we are taking a look at the weather data in Boston. The Wunderground Boston weather data contains daily average temperature, wind speed, humidity,

dewpoint, precipitation, and pressure from 2018 to 2020. We have hyperlinked the sources for the Bluebikes and Wunderground data sets.

Data Wrangling

Total rentals per day, *start2*, was used for simpler data visualization and initial data exploration. After combining and finding totals for all three datasets, we graphed each quantitative predictor and to see if predictors were normally distributed. Ultimately, we decided to log transform the frequency column in *start3.t* and *start4.t* due to the right skew. Temperature and wind speed were square-root and log adjusted respectively to correct for normality. Due to the perceived quantity of rain being less quantitative and more categorical, we adjusted our precipitation column to a categorical variable with no rain, low rain, and high rain, the latter two separated based on the average daily rainfall in Boston. Further trimming of the data was done by filtering out all stations which currently contain 0 docks (as they will be closed and not in service this year). This would result in inaccurate data for the docks variable or a changing value between years. Unfortunately, data of the original number of stations is inaccessible and not provided by Bluebikes, so we opted to remove these points. We also excluded all data points from the district “Everett” since Bluebikes only opened bike stations in this municipality in 2020, so we would not have adequate points of comparison from before the pandemic. We have included a list of all of our predictors in Table 1 as well as graphs of their transformations in the Appendix.

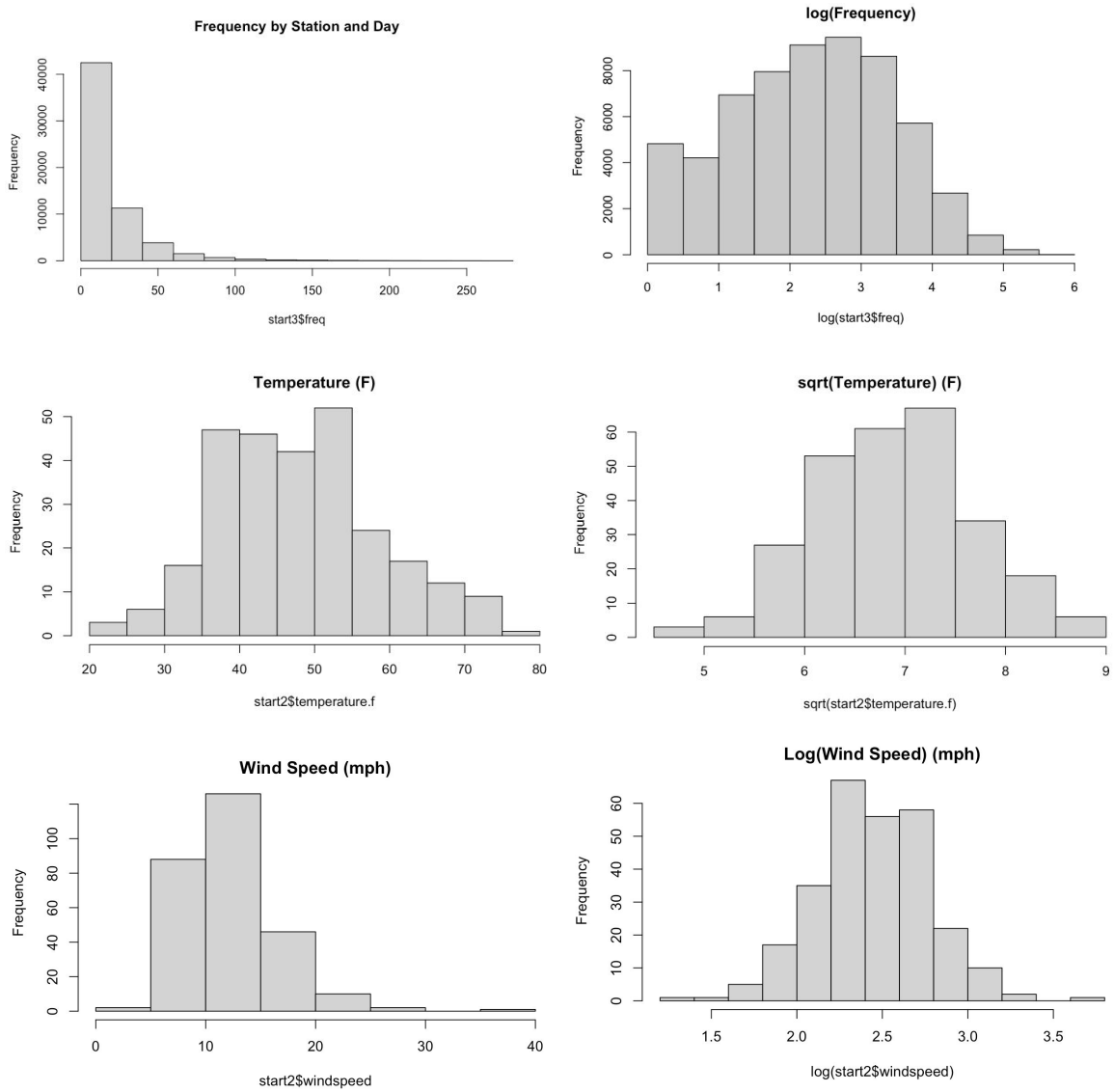

Figure 1. Histograms depicting untransformed (left) and transformed (right) predictor variables.

We also looked into possible sources of collinearity between variables, and noticed that dewpoint is highly correlated with temperature and humidity. This makes sense, as dewpoint is usually defined as a function of air temperature and relative humidity.

	temperature	dewpoint	humidity	windspeed	pressure
temperature	1.00	0.81	0.12	-0.14	-0.10
dewpoint	0.81	1.00	0.68	-0.16	-0.19
humidity	0.12	0.68	1.00	-0.07	-0.21
windspeed	-0.14	-0.16	-0.07	1.00	-0.35
pressure	-0.10	-0.19	-0.21	-0.35	1.00

Table 1. Correlation of weather variables.

Data Exploration

We chose to take a look at the relationships between frequency versus temperature, month, year, and weekend to determine possible confounders. It quickly became apparent that the transformed temperature and frequency variables have a significant relationship. The linear regression followed the equation of $\hat{y} = 15.303x - 44.666$ in which $\hat{y} =$ ***frequency*** and $x =$ ***temperature***. Both intercept and coefficient are significant as the *p*-value for both are well below 0.05. From the linear regression we can conclude that for every 1 degree increase in temperature we can expect the frequency of bike sharing to increase by about 15 riders.

We next investigate how bike trips change with the month. If we consider the data with all the years (2018, 2019, and 2020), the effect of months appears to be insignificant as the *p*-value of May is greater than 0.05. This motivates our separation of both months and years. As we can see below, the months of April and May usually see an increase in bikeshare frequency, but it decreases dramatically in 2020. Interestingly, bike sharing was higher in March 2020 than March 2018 or 2019, indicating bright coming months for Bluebikes. However, the lockdown orders in late March and Phase 1 reopening in May form a U-shaped curve for 2020.

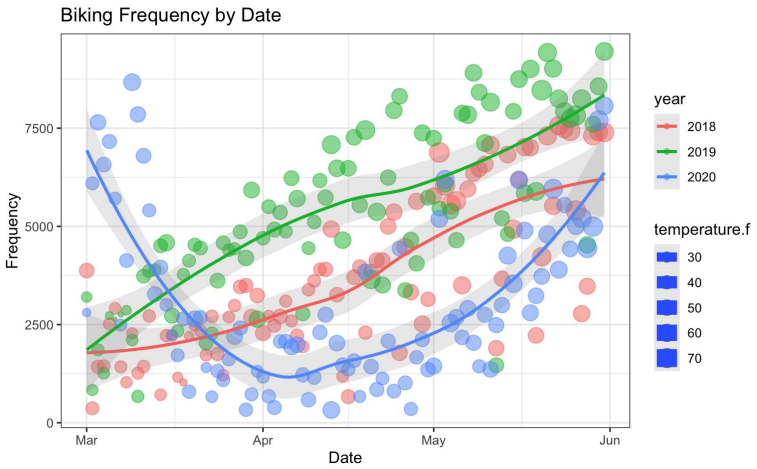

Figure 2. Scatterplot of biking frequency by date.

In addition to the variables considered above, we now take a look at the effect of weekends on bike sharing for 2018 and 2019. There is an 11.1 point drop in the transformed variable of bike trips per day on average over the weekend. This makes sense if more people bike to work on the weekdays whereas less people bike for leisure on the weekends. However in 2020 there is a 6.7 point increase in the transformed bike trips variable per day on average over the weekends. Interestingly it seems as though during quarantine more people are biking during the weekends. This makes sense since most people likely no longer have to bike to work due to lockdown orders and will only choose to bike for leisure during the weekends when they have the time. We have graphed the effect of weekends on bike sharing below.

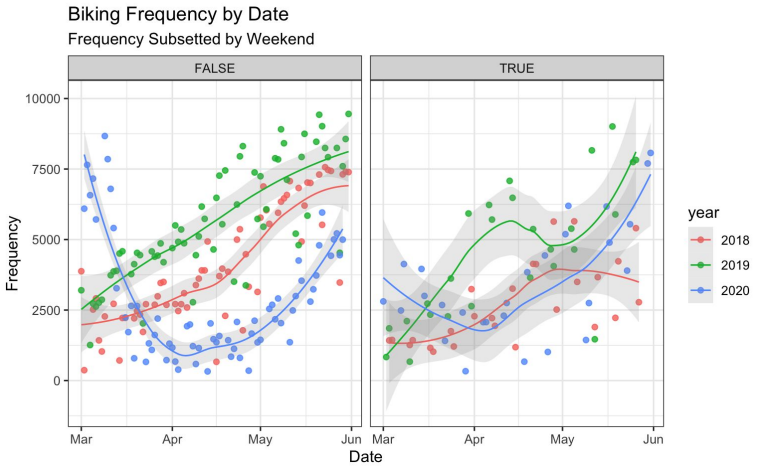

Figure 3. Scatterplot of biking frequency by date and weekend.

Both graphs reflect our findings from the linear models. It seems as though before the pandemic, most people used bikeshare during the weekdays, possibly to commute to work. However after the pandemic more people are biking during the weekend than the weekdays, as we can see by the 2020 trendline which is fairly close to the 2018 and 2019 lines in the weekend (TRUE) plot, instead of significantly below as on the weekday (FALSE) plot.

The plot and model output for the linear regressions are available in the Appendix.

3. Methods

Predictive Model

Our first hypothesis is that COVID-19 did have an impact on bike share, so we are looking at the effects of various variables on log bike share frequency (log.freq). By testing predictive models, we can determine more precisely what factors in addition to lockdown policy have the largest effects on bikeshare frequency.

I. OLS and Lasso

Our goal is to estimate the impact of COVID through the year by holding outside effects, such as weather, as constant as possible and determining possibly impactful interaction terms. Based on the exploratory data analysis, as we noted, due to high correlation of dewpoint with other variables, we omitted it in our OLS models for the predictive section. Our first linear model, fullmodel, contains no interaction terms. As we found in the exploratory data section, weekend and month have very different effects over different years, so it makes sense to have an interaction term for year * (weekend + month). Thus, our second linear model, model.int, only includes this as an additional interaction term. We also hypothesize that the impact on bikeshare usage may vary across districts, so we have two more variables with district as an additional variable: model.dist and model.moredist, where model.dist only includes interaction terms between district and time variables to investigate the impact of COVID by district and model.moredist investigates interaction terms of district with all variables. Finally, we weighted all of our linear models by the number of docks in order to reduce variance caused by the maximum number of bikes available per station.

For LASSO, we used both model.dist and model.moredist as baseline models. We decided to use the LASSO method to select the most important predictors because both linear models had a mix of many strong and moderate predictors, whereas Ridge regression is better suited for many weak to moderately strong variables and Stepwise variable selection is better for only a few strong predictors.

When we checked assumptions for this model, we saw that the model has somewhat reasonable residuals, and though the data points fan out moderately, the variance seems to stabilize. Also, there are some minor deviations from normality at the tails of the QQplot but we still believe it holds our assumption of normality.

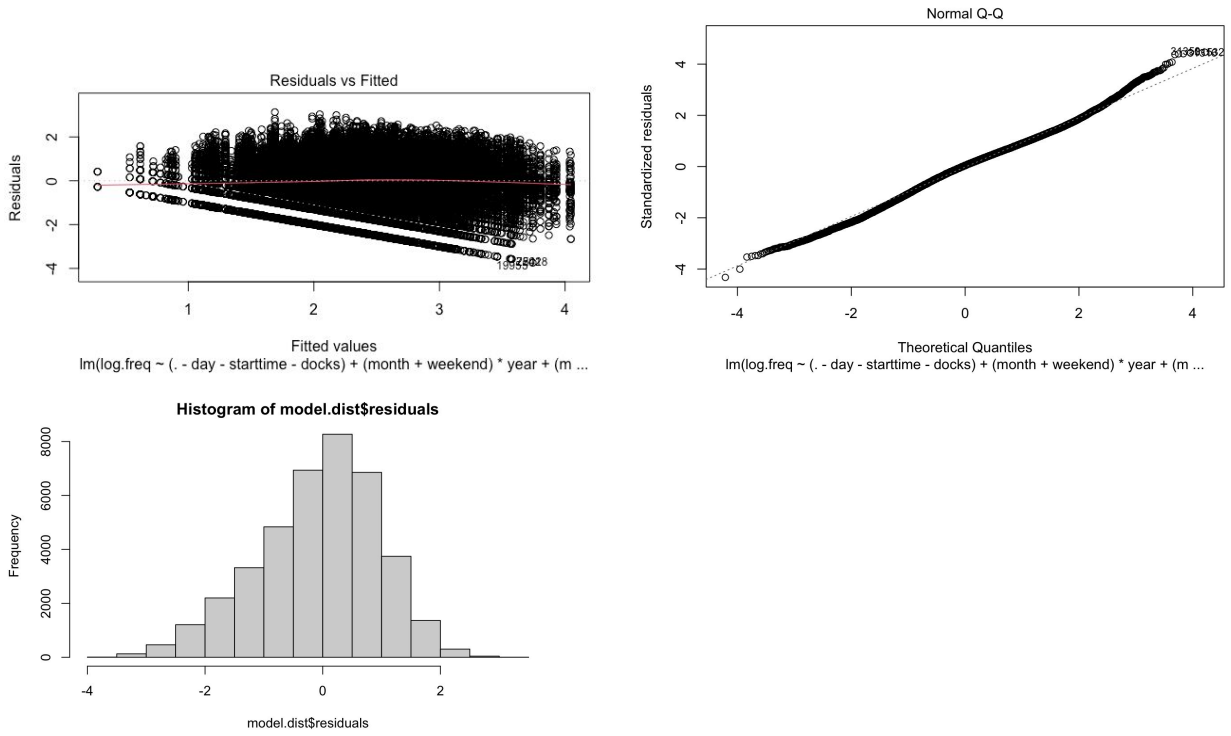

Figure 4. Assumption checking plots of OLS model.

II. Random Forest

Due to the likely multicollinearity between some predictors and possible interaction effects, it is likely that using all our predictors may not be the best model. We chose to further investigate using a random forest model as we saw that some relationships did not appear to be truly linear. Using a random forest will help us determine better predictive models. For this, we split our data into 80-20 train and test sets to help us tune random forest parameters. We considered the possibility of 1 to 10 different predictors (we have 10 total predictors) using 2, 4, 6, 8, 10, 15, and 20 maxnodes. We decided to run our random forest model on *start4.t* with the district daily totals in order to carefully investigate which effects of weather and other factors outside of the extreme variation between stations within a district. The aggregated effects of predictors on our district daily totals would better allow us to select the most significant predictors. Comparatively, *start3.t* with station daily totals would have over 60000 data points and large amounts of variation per station, resulting in possibly more nodes in our tree and more difficulty in determining significant relationships.

III. Mixed Effects Modeling

As we stated in our hypotheses, we are interested in investigating the differences in how the pandemic has affected Bluebikes bike share usage across different districts of Boston. Thus, we constructed some mixed effects models in order to introduce a more hierarchical structure for our data based on district, especially because there may be collinearity between variables based on whether bikers are commuters or leisure bikers, and Boston bikers may be more likely

to be bike commuters. Because districts all have vastly different numbers of total bike station docks (such as Brookline having only 113 total Bluebikes docks while Boston has over 3000), we decided to cluster based on district to help alleviate issues of potentially overfitting smaller districts.

	Boston	Brookline	Cambridge	Somerville
Bike station docks	3289	113	1338	425

Table 2. Number of bike station docks per district.

Initially, we attempted to fit a mixed effects model that was analogous to the OLS model we derived in part I by clustering by both year and district. As before, we weighted our data by docks in order to decrease variance caused by the maximum number of bikes possible. We decided to keep the predictors for weather as group-level fixed predictors, because one would estimate that the effects of weather would be the same across the entire Boston area, and we wanted to hold weather effects constant. Unfortunately there were too many predictors and the model was singular and also did not converge. This indicated that we needed to reduce the complexity of predictors that we were using, so we decided to focus on introducing random effects to only the variable Year, because we have selected our data such that all data points with the year as 2020 occur during the pandemic.

Hence we attempted to fit various mixed effects models where we clustered based on our four districts: a model where only the intercept was random, a model where only the coefficient of year was random, and a model where we had both fixed and random intercepts and coefficients. Unfortunately, even with reduced interaction terms and conditioning, rescaling individual variables, and adjusting the parameters of the control sequence, none of these models converged except for the model with only random intercepts, perhaps due to the complexity of our remaining predictors. Even so, our mixed models all indicated that all of the variables were significant, so we decided not to drop further predictors in order to be able to adequately compare our LMER model to our OLS model based on similar predictors.

Lockdown Fatigue Over Time: Methods

I. Linear and Quadratic OLS Models

Our third hypothesis is whether bikeshare data reflects the phenomenon of “lockdown fatigue”. After months of remaining cooped up inside, are Boston residents willing to bypass safety procedures and go outside? We attempt to quantitatively measure the effect of lockdown fatigue by determining whether there is an increase in bikeshare usership correlated to days since the lockdown began, while holding constant other non-pandemic related factors that may cause an increase in bike usership (such as increase in temperature, decrease in precipitation, et cetera).

To measure lockdown fatigue, we add the predictor *cov19* that measures days in lockdown. According to the Boston city official website, a state of emergency was declared on March 10th. So the *cov19* variable measures days since March 10th.

Our selection of dates from the beginning of March to the end of May in this dataset also minimizes the difference in bikeshare usage caused by phase 1 reopening, which was first introduced on May 26th and took even longer to implement. We choose not to model the cov19 variable with year and month due to likely collinearity. We add weather variables in order to hold weather factors constant and avoid confounding effects, and we weighted by docks in order to limit variance.

To test this hypothesis, we investigate whether a quadratic model is significantly better than a linear model with respect to the 'cov19' variable. If an upward facing quadratic model is more significant than a linear model, then this indicates that there may be some statistically significant upturn in bikeshare frequency correlated to days in lockdown, i.e. at the beginning of lockdown there is a drop in bike riding, and a gradual rise in bike share over time that may be caused by lockdown fatigue. The linear and quadratic model equations are shown in the Appendix.

Checking to see if our models violate any OLS assumptions, we see that our quadratic model fits normality fairly well. Also, looking at the residuals, we can see that our regression model may have some slight changes in variance, but variance becomes more consistent.

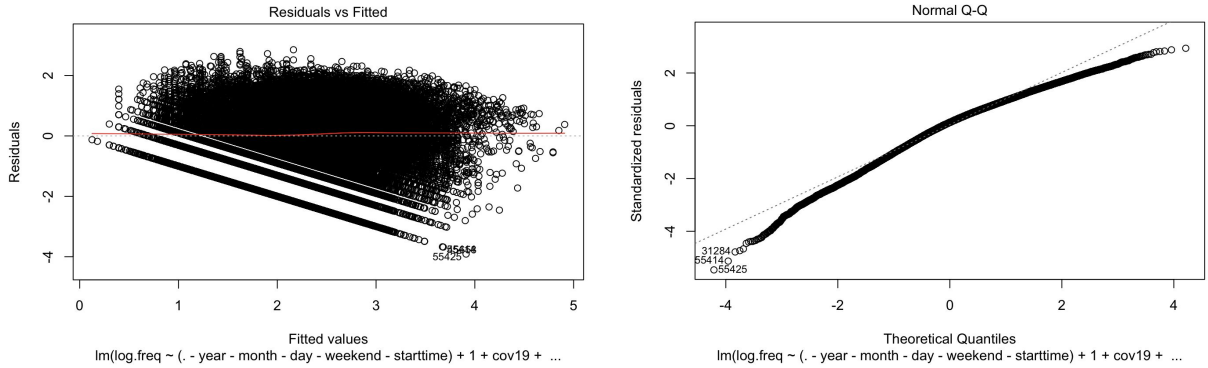

Figure 5. Assumption checking plots of linear and quadratic OLS models.

II. Mixed Effects Modeling

In addition to looking into whether there is lockdown fatigue at large, we are interested in investigating the differences in lockdown fatigue between districts. As some of the less urban districts are likely home to more people who are capable of working from home and live in larger houses and Boston may be home to more service workers or other essential workers who can reasonably commute around the city by bike, it makes sense that different districts may have people who are more eager to return to work (or get out of the house). Like in our earlier mixed models, we hope that we can introduce a hierarchical model to address collinearity between district clusters, and this can also help lessen possible overfitting from our models in part I.

We attempted to fit a mixed effects model that was analogous to the quadratic model in part I by clustering by district. While we had some difficulties in getting our models to converge, we were able to do so by rescaling all of our continuous predictors, setting our polynomial argument to raw=F, and making sure that variables that might have collinearity with cov19 (time-based variables, like year, month, and weekend) were removed. As before, we weighted our data by docks in order to decrease variance caused by the maximum number of bikes possible.

We attempted to fit two mixed effects models where we clustered based on our four districts: a model where the quadratic terms are fixed and only the intercept was random and a model where we had both fixed and random intercepts and coefficients for the polynomial terms. Because of how large our dataset is (roughly 40k data points in our training data) and how few districts there are, we are less concerned about overfitting by conditioning on districts. All of our predictors had significant coefficients.

4. Results

Predictive Model

We found that out of our predictive models, the LASSO model performed the best.

I. LASSO and OLS

All of our OLS models indicated that year:2020 was a significant predictor (with p-values under 0.05) and had negative coefficients, indicating that the average of log frequency in 2020 was significantly lower than in 2018.

We compared the likelihood ratios of our OLS models through a likelihood ratio test (asymptotic χ^2 test), AIC, BIC, and also computed the MSEs on the train and test sets. We have included our table of results below.

	df	AIC	BIC
fullmodel	15	113983.8	114138.4
model.int	20	117952.3	118124.1
model.dist	38	116154.9	116480.5
model.moredist	59	116172.4	116679.0

Table 3. AIC and BIC scores of OLS models.

Based on AIC and BIC, we found that the models model.dist and model.moredist using the districts as predictors both have fairly lower AIC and BIC compared to model.int (though *fullmodel* has the lowest AIC and BIC due to having the fewest degrees of freedom). Based on

our likelihood test, while we found that model.dist, model.moredist, and *fullmodel* had a statistically different fit from model.int.

	Train MSE	Test MSE	Difference
fullmodel	1.0529	1.0351	0.0178
model.dist	1.0492	1.0202	0.0290
model.moredist	1.0493	1.0202	0.0291

Table 4. Train and Test MSE of OLS models.

We found in this case that the most predictive model based on MSEs was model.dist, but the ANOVA results indicate that there is no statistically significant difference from model.moredist or fullmodel, so this result may vary based on the random sampling of our train and test data. Nonetheless, it is of note that none of these models seem to significantly overfit our data-- in fact, they all seem to perform slightly better on the test set than the training set.

We ultimately decided to choose model.dist for the purposes of this project, because even though all of the models were similar, model.dist has fewer interaction terms and is slightly more interpretable.

We find that the interaction terms here, (month + weekend) * year and (month + weekend + year) * district, support findings from our data exploration: that the usage of bikeshare on the weekend and by month varies in 2020, as evidenced by the significance of all of the following interaction terms:

Interaction Terms	Estimate	*p*-value
year2019:month04	0.1432	4.83e-05 ***
year2020:month04	- 0.6709	< 2e-16 ***
year2019:month05	0.1495	1.77e-05 ***
year2020:month05	- 0.3212	< 2e-16 ***
year2019:weekendTRUE	0.0784	0.0128 *
year2020:weekendTRUE	0.6777	< 2e-16 ***

Note: Significance levels are denoted by 0, '***': 0.001, '**': 0.01, '*': 0.05, '.': 0.1, ' ': 1

Table 5. OLS model interaction terms with year.

One possible interpretation for the significance of the interaction terms with the year variable is that the pandemic has shifted the demographic of users of bikeshare from work week commuters to those who bike for leisure over the weekend; similarly, the lockdown measures in spring, when bikeshare usage normally increases, likely caused a significant dip in bikeshare usage. The dramatically negative coefficients for the interaction terms with year2020 in particular illustrate how the mean of log frequency shifts down by roughly -0.67 and -0.32 points, respectively, in April and May during the pandemic.

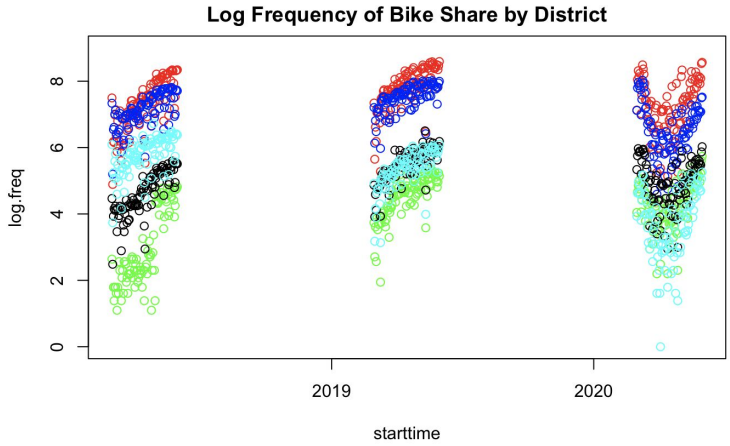

Figure 6. Log frequency of bike share by district.

Though not all of the interaction terms with district were significant, an interpretation for the interaction terms of year with district can be illustrated by the plot above, where we group points by district: it appears that the 2020 pandemic causes different magnitudes of change between districts. For example, Cambridge and Somerville have a more dramatic drop in 2020 than Boston (possibly due to the evacuation of students in the area), so it makes sense that the coefficient for the interaction terms *year2020:districtCambridge* and *year2020:districtSomerville* are negative as seen in Table xx , as this represents a larger drop in mean log frequency of bike riding during the pandemic.

Similarly, the increase of bikeshare usage over the months seems to follow different relationships between districts, particularly Cambridge, which we can observe through the interaction terms *month03:districtCambridge*, *month04:districtCambridge*, and *month05:districtCambridge*.

For LASSO, we used both model.dist and model.moredist as baseline models. The LASSO model model.dist dropped the variables *month03:districtBrookline*, *month03:districtSomerville*, and *year2020:districtBrookline*. The LASSO model model.moredist dropped *year2020:districtCambridge, districtBrookline:sqrt.temp, districtBrookline:humidity, districtCambridge:humidity, districtSomerville:humidity, districtBrookline:pressure, districtSomerville:pressure, districtCambridge:precipNone, and districtSomerville:precipNone.* Notably, model.moredist only dropped interaction terms rather than the main predictors.

Comparing the MSEs of these LASSO models, we find that they did indeed drop predictors that may have contributed to overfitting, as they outperform the simple OLS models.

Baseline LASSO	Train MSE	Test MSE	Difference
model.dist	1.0482	1.0168	0.0313
model.moredist	1.0481	1.0173	0.0308

Table 6. Baseline LASSO Train and Test MSE

II. Random Forest

Our best tree model had a maxnodes of 10 and considered 7 different predictors at each split. We see that our best tree had a higher RMSE on the training set than the test set, but this is observed for all other combinations of maxnodes and mtry. Ultimately, our model appears to explain for about 83% of the variance in the training data, and does a decent job in explaining about 81% of the variance in the test data set.

When we investigate the variable importance plot of the random forest, we see that our three most significant predictors were district, sqrt.temp, and year. District makes sense as the most significant predictor as more urban and rural areas will likely have different numbers of Bluebikes stations and accessibility. Temperature also makes sense as a significant predictor as hot and cold weather largely influences how many people are willing to bike. We see that year also appears to be a more relatively significant predictor of log.freq.

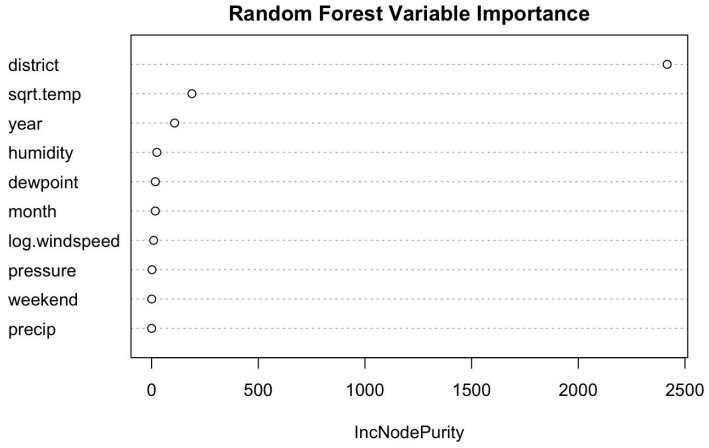

Figure 7. Random forest variable importance.

Plotting out log.freq against our most important quantitative predictor, sqrt.temp, below allows us to visualize the random forest. We are able to see the predictions based on many points in the dataset. Below, we have color coded the data points based on their district to highlight the

relative importance of the district as a categorical predictor of log.freq. It appears that there is a very slight positive relationship between sqrt.temp and log.freq. This makes sense as people are more likely to not bike when it is cold.

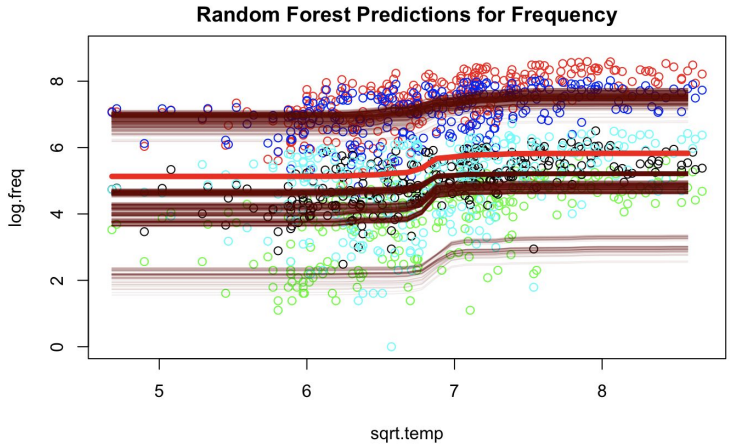

Figure 8. Random forest frequency predictions.

III. Mixed Effects Modeling

We investigate the effect of the year conditioned on the district for 3 variations of mixed effects models. Only the random intercept, fixed slope converged, so our results will only discuss model comparisons with that model.

We evaluated our mixed effect models through AIC and ANOVA.

	npar	AIC	BIC	logLik	deviance	Chi sq.	df	*p*-value
lmer2.2	16	114005	114142	-56987	113973			
model.dist	38	116123	116449	-58023	116047	0	22	1

Table 7. Correlation plot of weather variables.

Our AIC comparisons tell us that lmer2.2 (random intercept, fixed slope) is slightly more accurate than the OLS model.dist. However, the ANOVA comparison with model.dist linear model tells us that mixed effect model is not significantly better (p = 1) than the linear model. The very low intra-class correlation (ICC) for our mixed model is estimated to be 0.005, which is quite weak. The districts are not a large source of variability, which is consistent with our findings that our mixed effects model is not significant compared to lm model.dist. So, we can conclude that our lm model.dist and LASSO models are the most predictive.

Since the mixed effects models are not significant compared to any of the OLS models, this means that *we do not have significant evidence for our hypothesis that the effect of COVID on bikeshare is different when clustered across districts in greater Boston.*

Lockdown Fatigue Models

Our results from the pandemic fatigue models are:

I. Linear and Quadratic OLS Models

To compare our linear and quadratic OLS models, we compared the MSEs and performed an ESS F-test:

	Train MSE	Test MSE
linear	1.0323	1.0119
quadratic	0.9774	0.9610

Table 8. MSE for Linear and Quadratic OLS models.

	Res. Df	RSS	Df	Sum of Sq	F	*p*-value
linear	39570	708408				
quadratic	39571	751920	-1	-43512	2430.5	< 2.2e-16

Table 9. ANOVA (ESS F-test) for linear and quadratic OLS models.

We run MSE comparisons on the test and train data sets on the linear model. In both the train and test data sets, the quadratic model is better fit than the linear model. Since the train and test MSEs for the quadratic model are about the same, this suggests that the quadratic model is not overfit. So, we can conclude that our quadratic model is more predictive than our linear model. From our ANOVA comparison ($F = 2403.5$, $p < 2.2e\text{-}16$), the quadratic model is significantly better than the linear model at predicting bike ride frequency in Boston.

We now look at the summary plot for our quadratic model.

Interaction Terms	Estimate	*p*-value
districtBrookline	0.2187	2.95e-10 ***
districtCambridge	0.4240	< 2e-16 ***
districtSomerville	- 0.0892	2.19e-06 ***
...	...	...
cov19	- 0.0546	< 2e-16 ***
I(cov19^2)	0.0006	< 2e-16 ***

Note: Significance levels are denoted by 0, '***': 0.001, '**': 0.01, '*': 0.05, '.': 0.1, ' ': 1

Table 10. OLS model interaction terms.

From the summary plot, both terms in the quadratic equation with respect to day under lockdown are significant (p <2e-16 for both). So, the quadratic equation is a significant predictor of biking frequency. The portion of our quadratic equation with respect to the days under lockdown (cov19) is $\hat{y} = -5.258e-02 * x + 6.251e-04 * I(x^2) + ...$ where $\hat{y} =$ log bike share frequency and $x =$ *cov19* (days since COVID-19 lockdown). We omit other variables for clarity.

Notice that the quadratic model is upward facing since the second linear coefficient is negative (b = -5.258e-02 < 0) and the quadratic coefficient is positive (6.251e-04 > 0, with t-value 49.314). This means that with each increase of 6.251e-04 units of cov19^2, we have a one unit increase in log frequency (and thus each roughly 2x increase in frequency). Since the quadratic model to predict biking frequency is significant and U shaped with respect to days under lockdown, this suggests lockdown fatigue for bike riding during COVID. It also makes sense that the magnitude of this positive effect may appear small, as the cov19^2 variable goes up to a maximum of 82^2 and may hence require more extensive scaling down from its coefficient.

II. Mixed Effects Modeling

We evaluated our quadratic mixed effect models through AIC and ANOVA.

	npar	AIC	BIC	logLik	deviance	Chi sq.	df	*p*-value
lmerq.1	13	112363	112475	-56169	112337			
lmerq.3	16	112175	112312	-56071	112143	194.54	3	< 2.2e-16

Table 11. Likelihood ratio comparison between our two mixed effects models.

	npar	AIC	BIC	logLik	deviance	Chi sq.	df	*p*-value
quad	15	112470	112470	-56156	112312			
lmerq.3	16	1123175	112312	-56071	112143	169.09	1	< 2.2e-16

Table 12. Correlation plot of weather variables.

Looking at the AIC across all quadratic mixed effects models and the quadratic model, *lmerq.3 (random slope and random intercept) seems to be the most accurate model*. From our ANOVA comparison (χ^2_3 = 194.54, p < 2.2e-16), the quadratic days under lockdown effect on bikeshare is significant with the *random slopes* and intercepts model, while the *fixed slopes*/random intercept do not add extra explanatory power. Because our model includes both fixed and random coefficient terms, we see that the effect of days after lockdown on bikeshare frequency differs significantly from district to district.

Comparing the coefficients of our quadratic OLS with our mixed model, we can see that most of the coefficients are similar except for docks coefficient. The docks coefficient for the lmer model is roughly 5 times smaller than that of the quadratic lm model (0.05430 for the quadratic model and 2.537e-01 for the quadratic lmer model). This makes sense because in our quadratic mixed effects model, districts are no longer a fixed effect since we are clustering by districts. So in our quadratic lm model, the docks variable would have a greater impact as a fixed effect since the bikes per dock may also vary by district. Interestingly, our coefficient for the quadratic term is 5.388e+01, with *t*-value 10.220, much larger in magnitude than in our OLS quadratic model, where the quadratic term had coefficient 6.238e-04. This may suggest that there may still be confounders that we have not properly accounted for in our random effects model.

We can conclude that random intercept fixed slope mixed effect is the most predictive out of all the models. We now look at its summary plot.

	Estimate	*t* value
Intercept	2.4015	20.999
...	...	...
poly(cov19, deg=2, raw=F)1	- 59.6726	- 8.253
poly(cov19, deg=2, raw=F)2	53.8838	10.219

Table 13. Mixed model days since lockdown.

	Intercept	poly(cov19, deg = 2, raw = F)1	poly(cov19, deg = 2, raw = F)2
Boston	2.2657	-50.4185	49.6180
Brookline	2.4765	-47.0003	45.4038
Cambridge	2.6841	-76.5881	67.5367
Somerville	2.1795	-64.6835	52.9767

Table 14. Isolated effects by district.

The average effect of days under lockdown (*cov19*) quadratic model is $\hat{y} = -59.672584x + 53.883828x^2 + ...$ where $\hat{y} =$ bike share frequency and $x =$ *cov19* (days since COVID-19 lockdown). We omit other variables for clarity.

This is an upward facing, U-shaped parabola, since the second linear coefficient is negative ($b = -59.672584 < 0$) and the quadratic coefficient is positive ($a = 53.883828 > 0$). This average effect seems to vary slightly over districts. Surprisingly, Cambridge and Somerville have the steepest quadratic curves compared to Boston and Brookline when comparing the first derivatives. It seems that the lockdown fatigue seems to be slightly severe in Cambridge and Somerville than Boston. However, the effect on lockdown fatigue in Brookline is not noticeably different from Boston.

5. Conclusion and Discussion

Our initial hypothesis was that there was indeed an impact of COVID-19 on bike share patterns in the Boston metropolitan area. We saw that the year:2020 variable was significant across all OLS and LASSO models, hence indicating that there is a statistically significant negative impact during the year 2020. Our linear models also showed a significant shift in the distribution of bike share frequency: the demographic changed from mostly weekday riders in the pre-pandemic years to mostly weekend bike riders in 2020. This is an interesting shift that may reflect the prevalence of working from home due to the pandemic: commuters may no longer need to bike from work, and now may turn to Bluebikes for leisure instead. When focusing on COVID-19 impact between districts - the most significant predictor in our random forest model - we see a distinct cluster of data points by district. However, the mixed effects model did not converge so we are unable to fully gauge the influence of COVID-19 by district.

The second hypothesis we explored was about the effects of lockdown fatigue. We found that our quadratic, U-shaped, OLS model with respect to days under lockdown reflects an increase in bike sharing after an initial drop as the pandemic lockdown endured. When analyzing the effects of lockdown fatigue per district, our quadratic mixed effects models suggest that the effect of lockdown fatigue seems to be slightly more severe in Cambridge and Somerville than Boston, but lockdown fatigue is not noticeably different in Brookline compared to Boston.

There are several limitations to our model. Throughout data processing and cleaning, we noticed many instances where users checked out a bike for more than 24 hours. Our frequency count would only count them as one user, even though their bike **use** may have been over many days. Another limitation in our analysis was the simplification and loss of data due to Bluebikes being an expanding presence in Boston. As we mentioned in our exploratory data analysis section, we opted to remove stations from Everett and stations with incomplete docking data. Additionally, district population could be incorporated as a possible predictor variable, but due to the nature of ambiguous district borders and some docking stations residing between cities, it is difficult to have accurate prediction and usage counts (as the closest bike dock to a user technically living in Boston may be technically in Cambridge). Finally, while we were able to control for well-documented effects such as seasonal weather patterns, we were unable to fully factor in the full growth potential of Bluebikes had there not been any pandemic in 2020. As we did not have sufficient data to model the growth of Bluebikes, this fell beyond the scope of this paper; however, given that Bluebikes usage only increased from 2018 to 2019 and was even higher than 2019 in the days before pandemic procedure was announced in 2020, we can assume that the impact of Covid may have been even larger than we have determined.

We can see from our models that as the pandemic progresses, people do not really abide by stay at home orders. Of course, using Bluebikes may not be as insidious as directly violating social distancing guidelines, because it is possible to bike at a safe distance from others. However, it does shed light on the fact that it goes against human nature to stay isolated and because people are social creatures, and they will inherently seek interaction with other humans or nature. Perhaps a better government policy to ensure lockdown procedures are followed more closely is to incentivize people to stay at home by offering satisfactory compensation and

resources and that they need to consider COVID policy that encourages safe outdoor recreation rather than trying to enforce indoor isolation.

Taking a look at the impact of COVID-19 on bike share frequency has been enlightening because Bluebikes is such an important addition to Boston. However, this only offers a glimpse into the impact of COVID-19 and doesn't consider people that don't have the luxury to take bike rides for leisure. Bike sharing has a fee and thus we cannot see the effect that COVID-19 and lockdown fatigue has on the general population. For example, the shift in weekend bike riders is more reflective of a largely middle class demographic that can afford to stay at home, rather than people who may be working essential jobs or lower class workers who may have been laid off. In order to have a more comprehensive look at the scope of the pandemic, it would be interesting to expand to other modes of public transportation. Further exploration of bike sharing in other urban cities could also help us see the effects on a national scale. It could be worthwhile to investigate mean income per city and COVID policy as predictors to distinguish the regional demographic of bike share users.

Acknowledgements

We would like to thank the entire Stat 139 course staff for all their dedication and flexibility this semester, especially Sophie and Kevin for their advice during office hours. Thank you to Rohit for his timely emails and insight. This class has been a wonderful introduction to Linear Models. We would also like to thank Joe Blitzstein for preparing us for this class and all his support, and Neil Shepard for his wise words on truth, beauty, and goodness.

Appendix

EDA

For ease of readability, we have included the predictor variables that we use in the models as well as a brief description of the predictor variables, their variable type, and if a transformation occurred what type it was.

Table 15. Isolated effects by district.

Year
- The year the data was captured
- Factor (Integer)

Month
- The month the data was captured
- Factor (Integer)

Day
- The day of the month the data was captured
- Factor (Integer)

Start Time
- The date of the month the ride began
- Datetime Object

Start Station ID
- The station from which the bike was rented
- Factor (Integer)

Frequency
- The number of times a bike was rented that day
- Integer
- Transformation: log

Weekend Binary
- An indicator for whether a given day is a Saturday or Sunday (1 for weekend, 0 for weekday)
- Factor (Boolean)

District
- One of 5 municipalities in the Boston metropolitan area
- Categorical

Temperature (F)
- The average temperature of that day in Fahrenheit
- Double
- Transformation: square root

Dewpoint
- The dewpoint of that day
- Double
- Transformation: none

Humidity
- The average humidity that day
- Double
- Transformation: none

Windspeed
- The average wind speed that day
- Double
- Transformation: log

Pressure
- The air pressure that day
- Double
- Transformation: none

Precipitation
- The amount of precipitation that day in inches categorized by None, Low, High
- Factor
- Transformation: 3 level categorical

Docks
- The maximum number of bikes available to rent at each station
- Integer
- Transformation: none

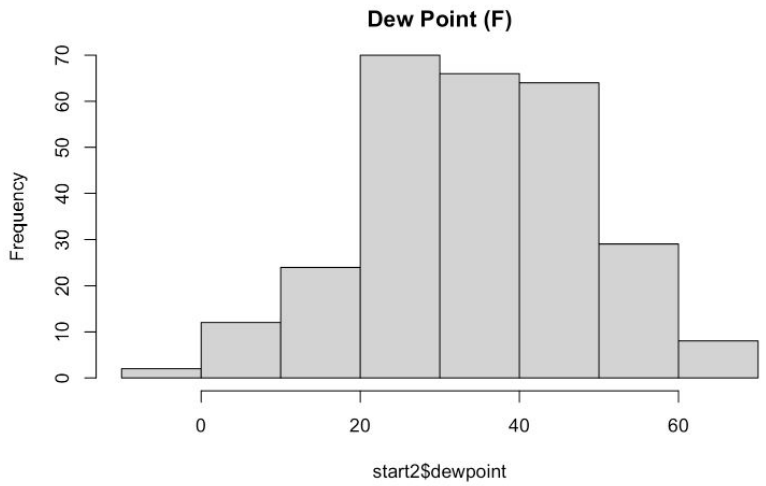

Figure 9. Histogram depicting untransformed average daily dew point (F).

We decided not to transform the dew point because it is somewhat symmetrical and applying any transformations results in a heavy right or left skew.

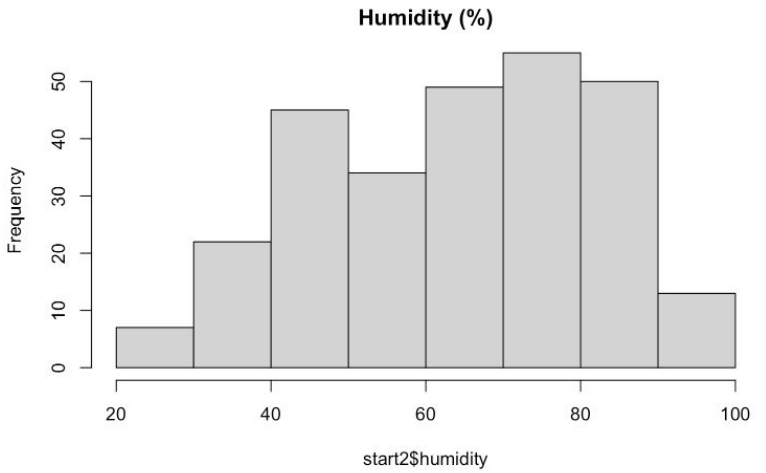

Figure 10. Histogram depicting untransformed average daily humidity (%).

We decided not to transform the humidity because it is somewhat symmetrical and applying any transformations results in a heavy right or left skew.

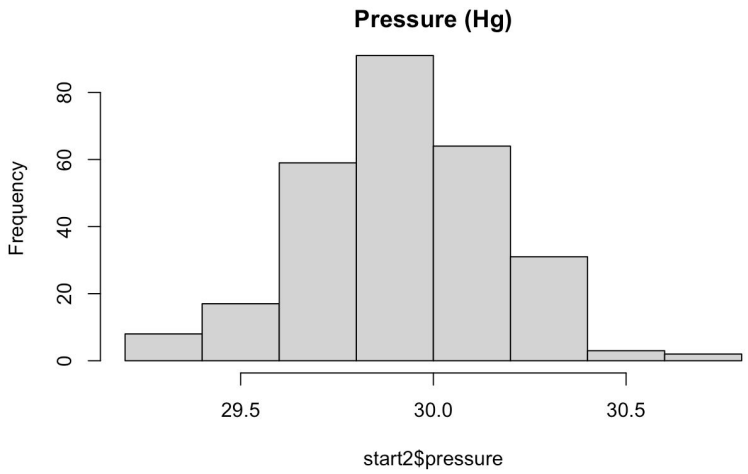

Figure 11. Histogram depicting untransformed average daily atmospheric pressure (Hg).

We decided not to transform the pressure variable because it is very symmetrical and normal-looking.

Below you can see the results of running a simple linear model comparing frequency and temperature. Temperature and frequency appear to have a significant relationship.

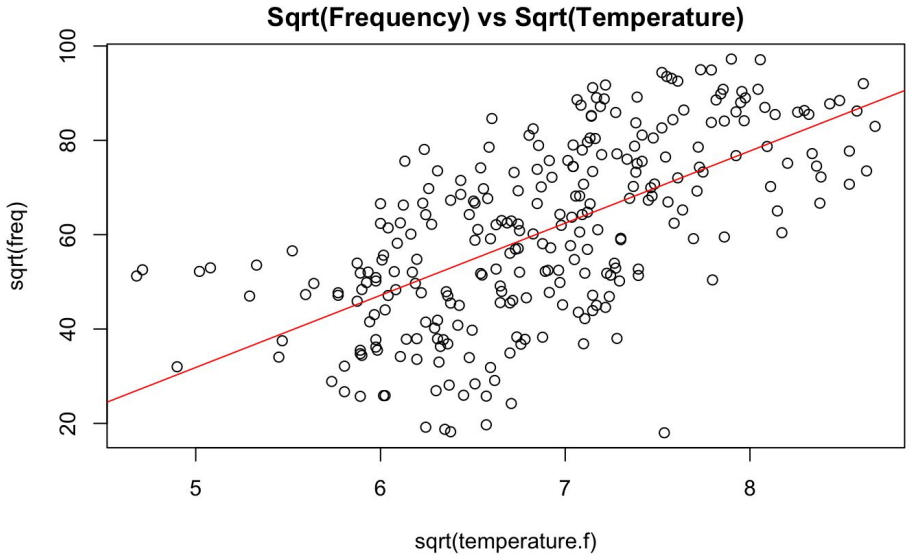

Figure 12. Scatterplot of sqrt(temperature) vs sqrt(frequency).

Model Coefficients:

```
lm(formula = sqrt(freq) ~ sqrt(temperature.f), data = start2)

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         -44.666      7.982  -5.596 5.34e-08 ***
```

```
sqrt(temperature.f)   15.303      1.153  13.277  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Full Model with Year (2018, 2019), Month, and Weekend

We consider the effect of weekends on bike share for 2018 and 2019. On average on weekends, there is a drop in 11.1 in the sqrt of the bike trips per day. This makes sense since more people bike to work on the weekdays, and may bike for leisure on the weekends.

Model Coefficients:

```
lm(formula = sqrt(freq) ~ sqrt(temperature.f) + humidity + log(windspeed) +
    month + weekend, data = start2[start2$year %in% c("2018",
    "2019"), ])

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          17.46767    9.06434   1.927 0.055581 .
sqrt(temperature.f)  12.87056    1.24597  10.330  < 2e-16 ***
humidity             -0.31509    0.03945  -7.988 1.73e-13 ***
log(windspeed)       -9.20700    1.99749  -4.609 7.74e-06 ***
month04               5.28363    1.90956   2.767 0.006263 **
month05               9.40448    2.61797   3.592 0.000425 ***
weekendTRUE         -11.12874    1.48557  -7.491 3.18e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Full Model with Year (2020), Month, and Weekend

We consider the effect of weekends on bike share in 2020. On average on weekends, there is an increase in 6.7 in the sqrt of the bike trips per day. Interestingly during quarantine, more people are biking during the weekends. This seems to be the case because people no longer have to bike to work due to lockdown orders and will likely only choose to bike for leisure during the weekends.

Model Coefficients:

```
lm(formula = sqrt(freq) ~ sqrt(temperature.f) + humidity + log(windspeed) +
    month + weekend, data = start2[start2$year %in% c("2020"),
    ])

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -13.03140   21.55123  -0.605  0.54701
sqrt(temperature.f)  17.49548    2.70589   6.466 6.08e-09 ***
humidity             -0.25087    0.07104  -3.531  0.00067 ***
log(windspeed)      -12.25800    4.52348  -2.710  0.00814 **
month04             -20.21888    3.21824  -6.283 1.36e-08 ***
month05             -13.01821    4.28091  -3.041  0.00313 **
weekendTRUE           6.69070    2.88911   2.316  0.02298 *
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Methods

Linear and Quadratic Model Equations

Linear Model with weather args	Quadratic Model, weather still constant
linear = lm(log.freq ~ (.-year -month -day -weekend -starttime) + **1+ cov19**, data=train, weights=docks)	quad = lm(log.freq ~ (.-year -month -day -weekend -starttime) + **1 + cov19 + I(cov19^2)**, data=train, weights=docks)

Results

ANOVA testing results

```
lmer2.2: log.freq ~ 1 + (. - day - starttime - district - docks) + (1 |
lmer2.2:     district)
model.dist: log.freq ~ (. - day - starttime - docks) + (month + weekend) *
model.dist:     year + (month + weekend + year) * district - district
           npar    AIC    BIC logLik deviance Chisq Df Pr(>Chisq)
lmer2.2      16 114005 114142 -56987   113973
model.dist   38 116123 116449 -58023   116047     0 22          1
```

```
lm(formula = log.freq ~ (. - year - month - day - weekend - starttime) +
    1 + cov19 + I(cov19^2), data = train, weights = docks)

Weighted Residuals:
     Min       1Q   Median       3Q      Max
-23.1322  -2.6688   0.5495   3.0071  12.4610

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -2.092e-01  7.949e-01  -0.263 0.792376
districtBrookline   2.187e-01  3.470e-02   6.303 2.95e-10 ***
districtCambridge   4.240e-01  1.118e-02  37.941  < 2e-16 ***
districtSomerville -8.917e-02  1.883e-02  -4.736 2.19e-06 ***
docks               5.430e-02  8.611e-04  63.065  < 2e-16 ***
sqrt.temp           1.564e-01  4.420e-02   3.539 0.000402 ***
dewpoint            2.045e-02  3.398e-03   6.019 1.77e-09 ***
humidity           -1.880e-02  1.577e-03 -11.926  < 2e-16 ***
log.windspeed      -4.191e-01  1.729e-02 -24.242  < 2e-16 ***
pressure            7.201e-02  2.430e-02   2.963 0.003045 **
```

```
precipLow          -3.583e-02  1.579e-02  -2.270 0.023228 *
precipNone         -8.286e-02  1.456e-02  -5.690 1.28e-08 ***
cov19              -5.258e-02  8.352e-04 -62.956  < 2e-16 ***
I(cov19^2)          6.251e-04  1.268e-05  49.300  < 2e-16 ***
```

ANOVA test results:

```
       npar    AIC    BIC logLik deviance  Chisq Df Pr(>Chisq)
lmerq.1   13 112363 112475 -56169   112337
lmerq.3   16 112175 112312 -56071   112143 194.54  3  < 2.2e-16 ***
```

```
Scaled residuals:
    Min      1Q  Median      3Q     Max
-5.4359 -0.6298  0.1295  0.7093  2.8558

Random effects:
 Groups     Name                           Variance  Std.Dev. Corr
 district   (Intercept)                      0.05146  0.2268
 district.1 poly(cov19, deg = 2, raw = F)1 196.67942 14.0242
            poly(cov19, deg = 2, raw = F)2 100.24863 10.0124  -0.91
 Residual                                   17.80655  4.2198
Number of obs: 39584, groups:  district, 4

Fixed effects:
                                  Estimate Std. Error t value
(Intercept)                       2.401517   0.114365  20.999
scale(docks)                      0.253690   0.004007  63.310
scale(sqrt.temp)                  0.118458   0.033566   3.529
scale(dewpoint)                   0.280663   0.046260   6.067
scale(humidity)                  -0.334035   0.027888 -11.978
scale(log.windspeed)             -0.134561   0.005534 -24.314
scale(pressure)                   0.017603   0.005944   2.961
precipLow                        -0.035341   0.015744  -2.245
precipNone                       -0.082224   0.014523  -5.662
poly(cov19, deg = 2, raw = F)1 -59.672584   7.230217  -8.253
poly(cov19, deg = 2, raw = F)2  53.883828   5.272941  10.219
```