Uber Travel Speeds in Seattle

Authors: Aboli Moroney, Harini Ram Prasad, Mayank Goel, Samarth Modi

TABLE OF CONTENTS

1. Abstract

Uber publicly provides the data about the hourly average speeds of Uber vehicles on the streets of major cities of the world. With a vision to understand the movement trends of Uber rides on the streets of Seattle, we explored the data and formalised three hypotheses around this data. We started by understanding the data profile and checking for assumptions in the same. We further evaluated the most appropriate tests like Analysis of Variance (ANOVA) test, Welch test and Z test to check the validity of our hypotheses. We interpreted the results and made inferences based on them. We concluded our analysis by extending the inferences obtained and discussing the limitations.

Keywords: *Uber, Hypothesis, Seattle, ANOVA, Welch test, Z-test*

2. Introduction

With the advent of ride hailing services, Uber arose as a leading service provider over years. These services have received a wide acceptance, and today it has become an almost inevitable part of our daily commute needs. Given this wide network, my group was interested in exploring and understanding movement trends of Uber rides on the streets of Seattle city.

We used "Uber Movement Speeds" dataset to formulate our hypotheses to explore the following areas of interest: 1. Seasonality in the average speed by day of week 2. Impact of popular festivals on the average speed, and 3. Average Speed during different peak hours of the day.

We believe that this analysis might be of interest to anyone curious to understand the Uber movement speed trends in Seattle, including office goers, students, Seattle city planning corporation, etc. The link to the github repository containing our entire analysis and data can be found here: https://github.com/mickkygoel/Data-557-Uber-Speed-Analysis

3. Dataset Description

3.1. Data Profile

Uber publicly provides the data about the speeds of Uber vehicles in the "Uber Movement Speeds" data, which contains aggregated speeds of Uber rides by street segments at hourly granularity for the years 2018 and 2019. The dataset contains more than 46 million data points per year and has a size of over 15 GB. However, Uber provides this data as a separate file for each month, taking up space of about 1.2GB - 1.5GB for each month. To process data of this size using the limited computational capacity in R, we had to batch process our data for each month, and finally collate data for the individual months to yield the complete dataset.

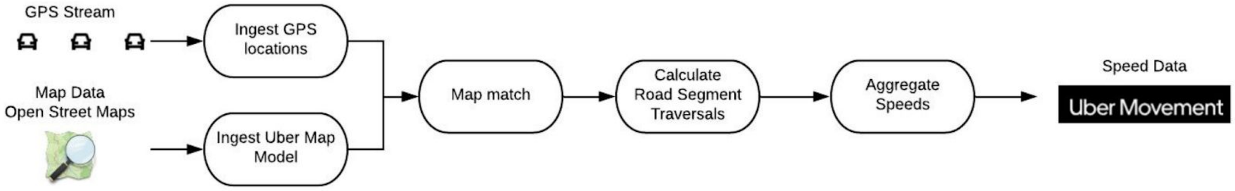


Fig. 1 - The Overview of Speed Data Calculation Methodology

Figure 1 illustrates the steps used in calculation of Movement speed data:

1. **Input data (GPS location and Map Data) ingestion:** Location data is collected from Uber Driver-partner app using the GPS and from a 3rd party map data provider, Open Street Map (OSM).
2. **Map matching:** Hidden Markov Chain Model to assign GPS location ping to corresponding OSM street with the highest probability.
3. **Traversal Speed Calculation:** Time difference between when the driver enters the street segment and when the driver leaves the segment is calculated, and Speed is consequently calculated as the distance of the segment divided by this time.
4. **Speed aggregation:** The traversals for each segment are aggregated into hourly time windows.

3.2. Data Independence and Randomness

Based on the data collection method discussed above, the data can be considered as a randomized sample of the entire population of Uber rides. The Uber movement trends dataset covers a high percentage of the population, and the randomness can be accounted due to the following reasons:

1. It contains street level data which is calculated from Uber vehicles of all types, be it UberX or Uber SUV or any other type of vehicle.
2. It contains hourly aggregated data, which is calculated from Uber vehicles moving through a street at any point of time, and at any frequency in an hour.
3. We haven't mapped OSM IDs to corresponding streets in seattle. Thus, the aggregated data contains a randomised order in streets, with least locality bias.
4. The varying environmental conditions like the weather conditions add to the randomization.
5. The wait time at traffic signals affects the average speed on the street, adding to the randomization.
6. Varying traffic conditions influencing vehicular speeds while travelling across a street also add to the randomization.
7. Speeds measured on one street are not dependent on the speeds measured on another speed.

Hence, we can assume that the samples as well as the observations within the sample are independent of each other.

3.3. Data Aggregation

Uber Movements Speeds provides hourly aggregated speeds of Uber vehicles by street segments, i.e. we have information about the average speed of Uber vehicles on every street under consideration in Seattle, for every hour of every day in 2018 and 2019. The start and end points of the streets are indicated by OSM start node ID and end node ID respectively.

Despite multiple trials to map OSM IDs to actual street names, we were unable to get this mapping. Hence any analysis at a street or area level was out of scope due to this missing mapping information. Thus, we aggregated the mean speeds on the streets by taking an average of mean speeds on all the streets, i.e now we have information about the average speed of Uber vehicles in Seattle for every hour of every day in 2018 and 2019. We have chosen to use this approximation due to the unavailability of relevant traffic metrics, some of which are:

1. Car types and counts per type for each street
2. Congestions during different times in a day
3. Type of street (highway vs. local street)

Further aggregation was performed as per the needs of each hypothesis.

3.4. Exploratory Data Analysis

Taking a look at the daily average speeds, we observe some periodic spikes, which might show some weekly trends. Also, we see that festive days have some deviation from the normal for example, Christmas has fairly high average speed and Valentines day has a fairly low average speed.

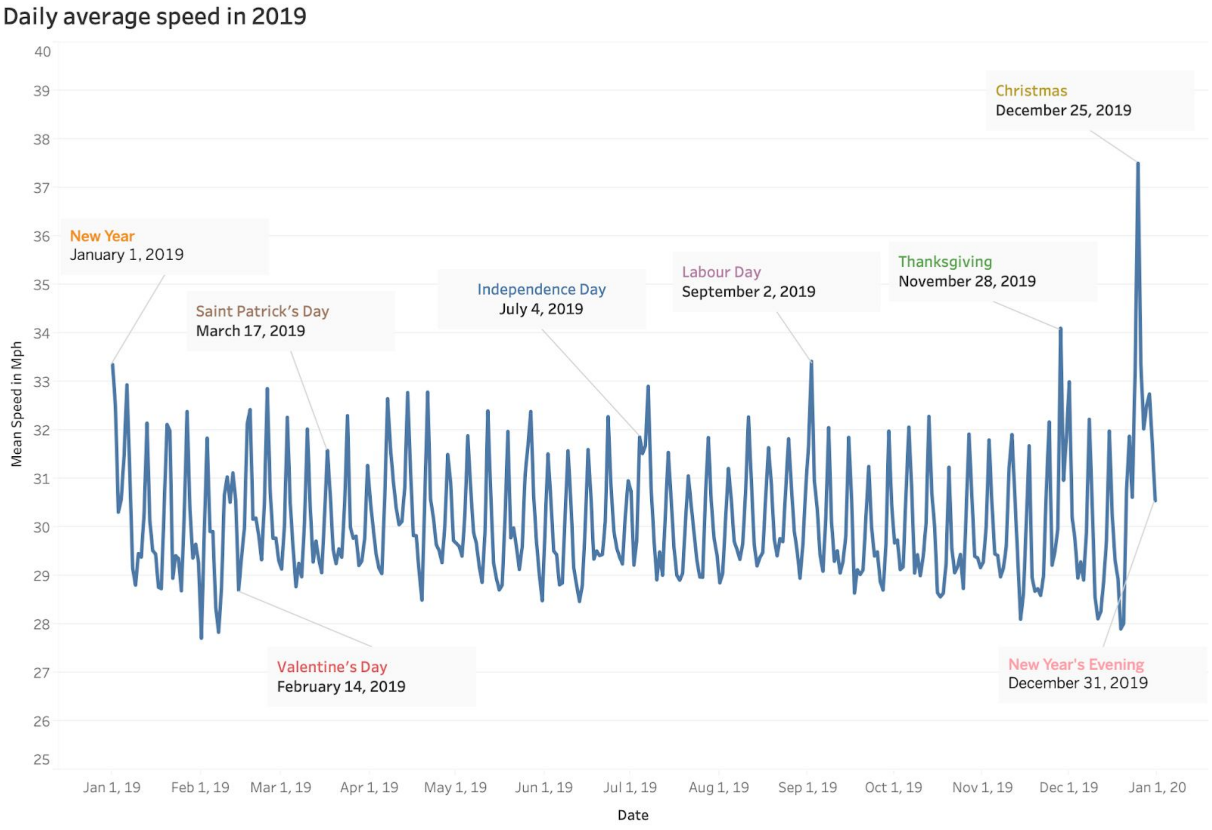


Fig. 2 - Daily aggregated mean travel speed in a year

4. Statistical Methods

4.1. Day of Week

We were interested in exploring how the mean travel speed differs with ‘day of the week’ in Seattle. Our way of life has a lot of factors which can influence how people behave on different days of the week. These differences can influence how people travel as well. For example, unwillingness to resume work on a Monday morning, better known as the ‘Monday Blues’ might affect how fast people drive on their way to work. Also, weekends are known to be less busy and lazy and people may not flock the roads as much. Similarly, many such reasons can pronounce the differences between speed on different days of the week. Through this analysis, we want to explore how the mean travel speed changes based on the day of the week and also see if there is any inherent seasonality or pattern which we observe.

4.1.1. Data Exploration

Exploratory Analysis

Before diving into the statistical tests, we were curious to explore the dataset to find useful trends and information about the data that could help us shape our analysis. We found some interesting patterns for mean speed over different days of the week.

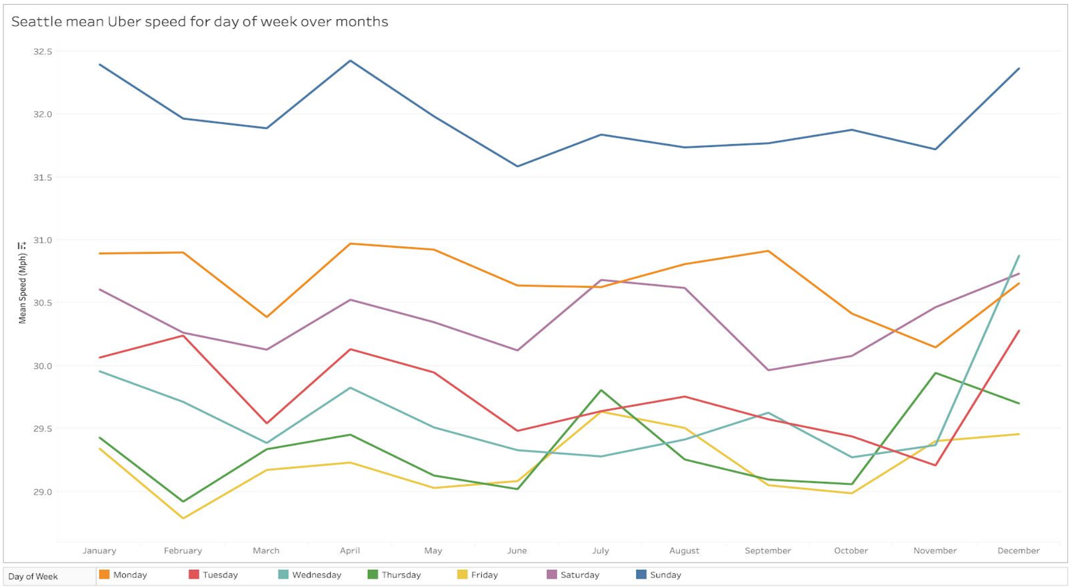


Fig. 3 - Seattle mean Uber speed for day of week over months in 2019

Note: Y-axis scale has been adjusted for the range to provide a magnified view of the trends

From the chart above, we can infer the following:

- Sundays are consistently faster than other days all around the year
- Mondays are surprisingly faster than Saturdays or any other weekday for most part of the year

- There are some big disruptions in mean speed in the month of December, possibly due to the festive times of Christmas and New Year as well as weather changes.

Data Preparation

For this particular question, we wanted to compare the daily averages for each day against each other. Hence, we aggregated the available street-level hourly data using 'average' to obtain the daily mean speeds in Seattle. Here is a snapshot of the prepared dataset for this hypothesis. As shown in table below, we have exactly one value for each day of the year 2019.

Year	Month	Day	Day of Week	Mean Speed (mph)
2019	1	1	Tuesday	33.35249
2019	1	2	Wednesday	32.47098
2019	1	3	Thursday	30.30966
...	...	...	...	...
2019	12	31	Tuesday	30.54495

Table 1 - Dataset snapshot for Hypothesis 1

4.1.2. Outlier Treatment

We observed that the dataset had some significant outliers aligning with some festivals or national holidays. For example, a particular Wednesday had a much higher mean speed compared to the rest of the Wednesdays, as this Wednesday happened to occur on 26th December (Boxing Day) during which most people are on holidays.

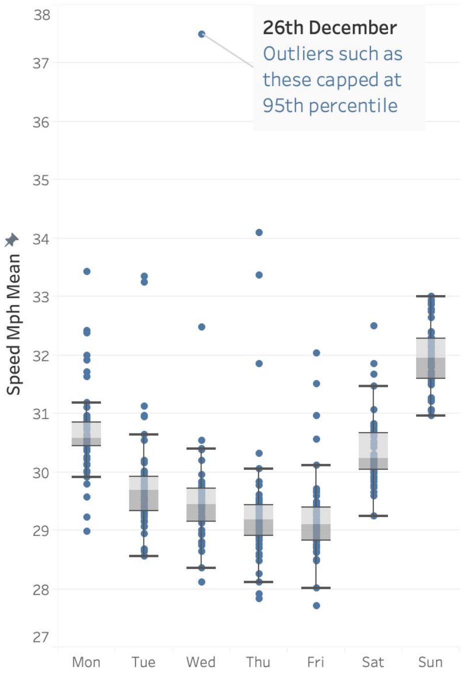


Fig. 4 - Uncapped mean speed

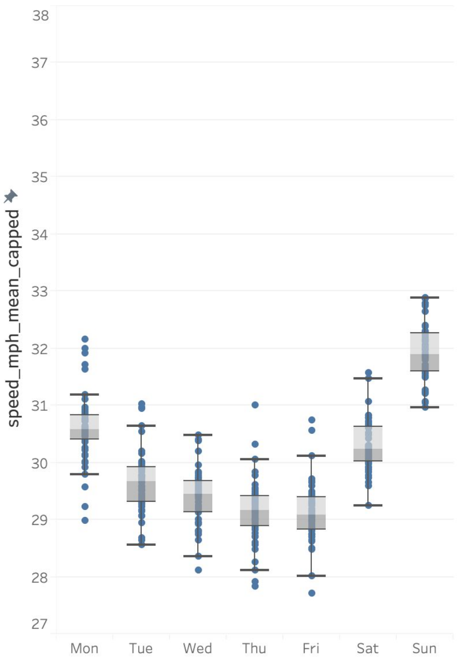


Fig. 5 - Capped mean speed

Since the underlying sentiment of our hypothesis is to compare mean speed for an average day with other days of the week, it made sense for us to cap such outliers. We prefered capping outliers instead of dropping the values as we were cautious about losing data, and also wanted to maintain an equivalent sample size for all days. We chose to cap the outliers at 95th percentile instead of the conventional 99th percentile as our sample size (51-52) was less than 100, and capping at 99th percentile did not remove even the most extreme outlier. We decided to cap only the upper bounds of the data points, as there was no major outlier from the lower bound.

An additional advantage of capping the outliers was that the variances across days of the week became much more consistent with each other. In the figures 4 and 5, we can see how the distribution of the data changes after the outlier treatment.

4.1.3. Assumption Validation

We further explored the distribution of the dataset to evaluate different assumptions that are met for this scenario. This was done to evaluate the most suitable statistical test for this hypothesis.

Normality

On plotting the histogram and q-q plot for the dataset, we inferred that the dataset is not exactly normal. However, none of the plots provided any significant evidence against normality. Therefore, we felt confident in assuming normal distribution for statistical testing.

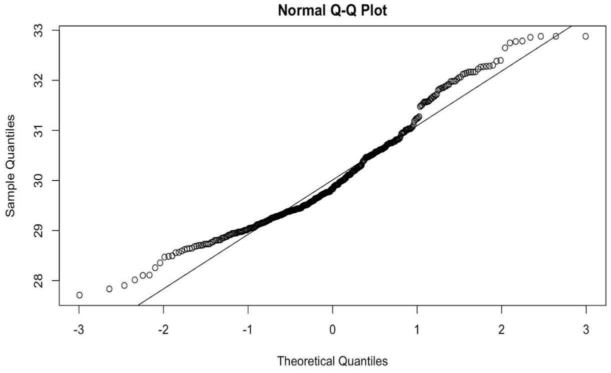


Fig. 6 - Histogram for normality test

Fig. 7 - Q-Q plot for normality test

Variance

The initial data without outlier capping showed some significant deviance in variances for each day of the week. However, after capping the outliers at the 95th percentile, the variances among days of the week become much smaller and consistent, and their range also reduces significantly. This is also evident from the figure 5 above. The table below compares the variances before and after the capping. Thus, we felt confident in assuming equality of variance.

We also tested for equality of variance through our residual, which provides much more robust proof in support of this assumption and is covered in the section 4.1.4 below.

Day of week	Variance (uncapped)	Variance (capped)
Monday	0.5808846	0.4467365
Tuesday	0.7862165	0.3533132
Wednesday	1.6810621	0.2717846
Thursday	1.1641402	0.4258321
Friday	0.5453930	0.3635415
Saturday	0.3365615	0.2547196
Sunday	0.2739123	0.2662607

Table 2 - Day of week - capped and uncapped variances

4.1.4. Hypothesis Testing

Null Hypothesis

The average travel speed for Uber rides was the same for all 7 days of the week in Seattle in 2019.

$$\mu_{mon} = \mu_{tue} = \mu_{wed} = \mu_{thu} = \mu_{fri} = \mu_{sat} = \mu_{sun}$$

where μ_{mon} through μ_{sun} are the average Uber travel speeds for Monday through Sunday respectively for Seattle in 2019.

Test Selection

As we had 7 different groups for each day of the week to compare and we met all assumptions for the ANOVA test (Table 3), ANOVA was the most logical and suitable choice for this test.

Assumption / Check	Status	Reference
Independence	**Pass**	Section 3.4 Exploratory Data Analysis
Equal variance	**Pass**	Section 4.1.2 Assumption Validation
Normality	**Pass**	Section 4.1.2 Assumption Validation

Table 3 - Check for assumptions for 'Day of Week' hypothesis

Test Results

ANOVA test results gave a high F-statistics value of 152.8, and a p-value less than <2e-16. Therefore, we were able to reject the null hypothesis that the average travel speed for Uber rides was the same for all 7 days of the week in Seattle in 2019.

```
                        Df      Sum Sq      Mean Sq      F value      Pr(>F)
factor(dayofweek)       6       312.2       52.04        152.8        <2e-16
Residuals          356     121.2       0.34
```

4.1.5. Residual Diagnostics

It was important for us to validate the correctness of our result. Therefore, we performed some residual diagnostics after fitting the ANOVA model on our dataset to test for our assumptions. We plotted the boxplot and histogram for the residual as shown below. The two plots infuse a lot more confidence in this test as we can infer the following:

1. The box-plot on the left shows no evidence against equality of variance, hence the equal variance assumption is justified.
2. The histogram on the right shows a fairly normal distribution of errors, hence normality assumption is justified.

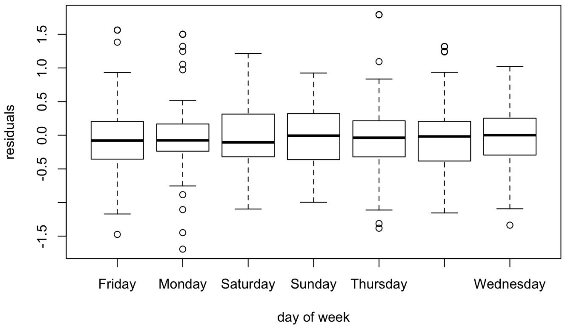


Fig. 8 - Box plot of residual for day of week

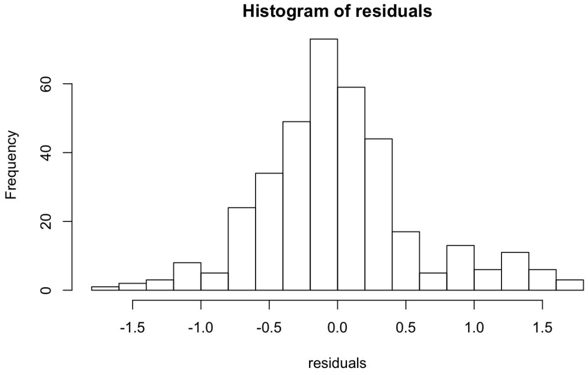


Fig. 9 - Histogram of residuals

4.1.6. Inference

We can infer that the mean speed for Uber rides in Seattle in 2019 is not the same for each day of the week. This is in agreement with our results from descriptive statistics and trends analysis. It is important to note that mean speed is taken across streets without accounting for the number of cars, i.e. each street is given equal weightage regardless of the number of cars passing through it.

To drill down further, we also performed an ANOVA test by grouping days as weekdays vs weekends. This test (F-stat: 229.3, p-value: < 2e-16) also gave us a similar conclusion that mean speed is not the same for different days of the week.

Possible Explanation: The result is intuitive and seems logical as traffic is highly influenced by work commute, which affects the mean speed on working days, but not on weekends. It can further be influenced by factors such as employees working from home, different office timings for Fridays, people going out on Saturday evenings etc.

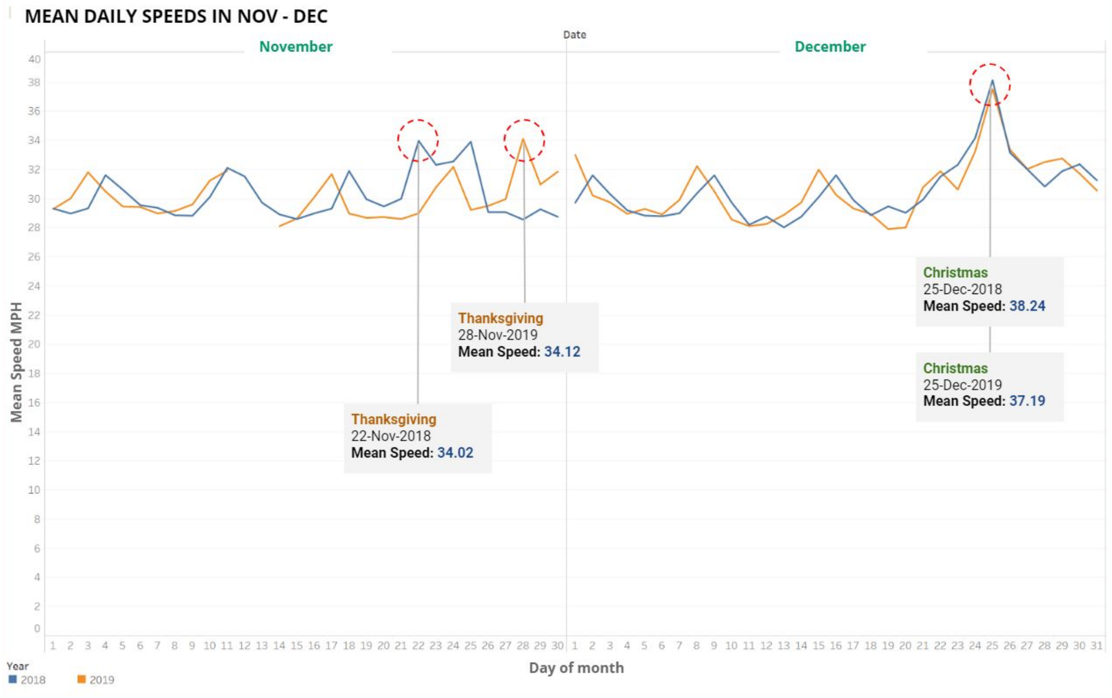
4.2. Festivals

We were interested in exploring how the mean travel speed would differ for different festivals in Seattle. We had noticed during our exploratory phase that festive days often showed a deviation in mean travel speeds from the expected trends (Refer 3.4 Fig-2). Hence, we decided to compare the 2 most popular festivals in Seattle which had shown the highest deviations - Thanksgiving and Christmas. Based on our knowledge and observation, both these festivals are mostly celebrated indoors which could explain the spikes we observe. We decided to compare these two festivals to check if there was any significant difference between them.

4.2.1. Data Exploration

Trend Analysis

Before performing statistical tests, we did some data exploration to better understand our festival data. The trend chart below contains daily data for November and December for the years 2018 and 2019. We notice spikes in mean speed for Uber rides around the days of Thanksgiving and Christmas in both the years.



Thanksgiving
22-Nov-2018
Mean Speed: 34.02

Thanksgiving
28-Nov-2019
Mean Speed: 34.12

Christmas
25-Dec-2018
Mean Speed: 38.24

Christmas
25-Dec-2019
Mean Speed: 37.19

Fig. 10 - Trends for November and December in 2018 and 2019

Data Preparation

We selected the day of the festival as well as consequent days to the weekend nearest to the festival. This selection was made keeping in mind that most of the people take time off to celebrate these popular festivals in conjunction with a weekend. Below are the actual date ranges that were filtered for this analysis from the data provided by Uber:

Thanksgiving: Sample of 10 days in total

- In 2018: Wed, 21 Nov to Sun, 25 Nov
- In 2019: Wed, 27 Nov to Sun, 01 Dec

Christmas: Sample of 11 days in total

- In 2018: Sat, 22 Dec to Wed, 26 Dec
- In 2019: Tue, 24 Dec to Sun, 29 Dec

The data which was originally at an hourly level by street was averaged to get mean speed per day across all streets. Below is a snapshot of the prepared data which was used to test the hypothesis.

Year	Month	Day	Festival	Mean Speed (mph)
2018	1	1	Thanksgiving	29.98616
2018	1	2	Thanksgiving	33.95708
...	...	...	...	...
2019	12	29	Christmas	32.74694

Table 4 - Snapshot of prepared data

4.2.2. Assumption Validation

We further studied the distribution of the data to evaluate the most suitable statistical test:

Statistic for Mean Speed	Min	1st Qu.	Median	**Mean (μ)**	3rd Qu.	Max	Std. Dev.	**Variance (σ^2)**	**Sample Size (n)**
Thanksgiving	29.96	31.19	32.43	**32.26**	33.68	34.1	1.56	**2.43**	**10**
Christmas	31.51	32.41	33.15	**33.69**	33.76	38.11	2.16	**4.66**	**11**

Table 5 - Descriptive statistics

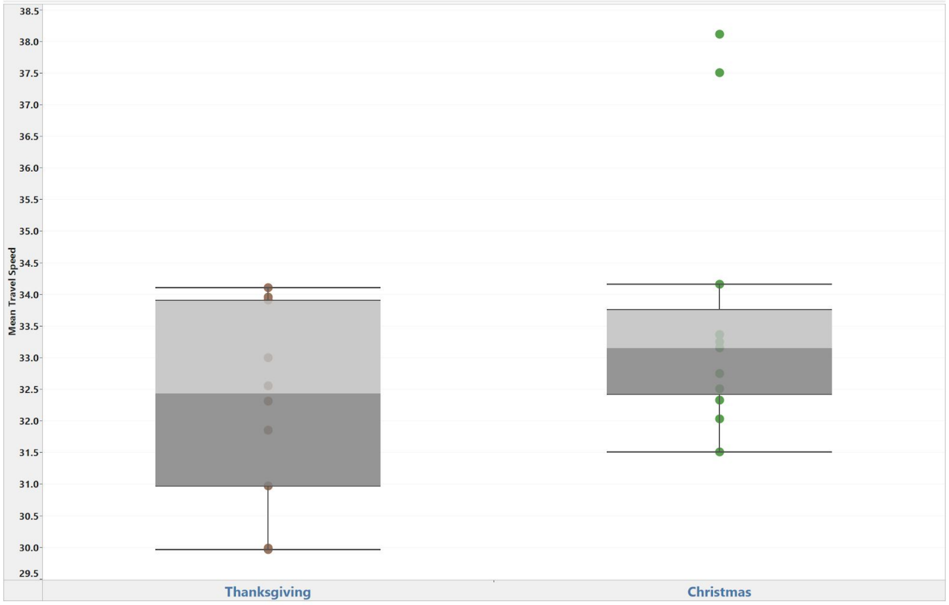


Fig. 11 - Box plots for Thanksgiving and Christmas mean speeds

From the descriptive statistics (Refer Table 5), we could derive that the mean speeds for the 2 groups were quite close to each other but the variances were different. Christmas days showed higher variances as compared to Thanksgiving and also had a few outliers (refer Fig11). The size of both the samples was quite small and comparable to each other.

Normality

On plotting Q-Q plot for the festival dataset, we could infer that the dataset follows a fairly normal distribution with the exception of a few outliers. As slight deviation from normal distribution is common, we felt confident in assuming that there was no strong evidence against normality for statistical testing.

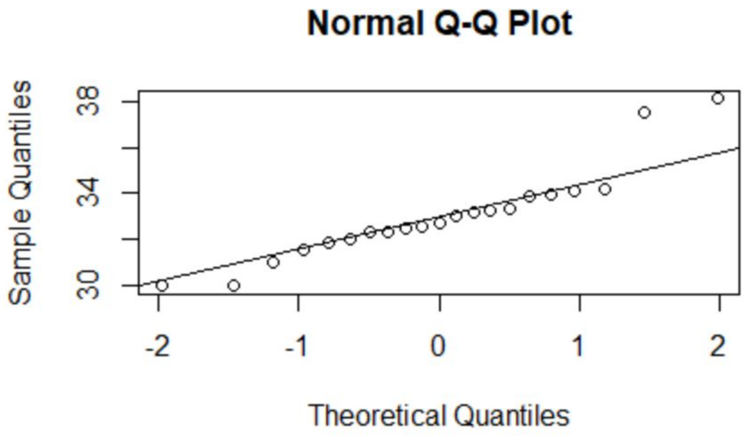


Fig. 12 - Q-Q plot to check normality of festival mean speeds data

4.2.3. Hypothesis Testing

Null Hypothesis

The average travel speed for Uber rides was the same for both festive periods around Christmas and Thanksgiving in Seattle in 2018 and 2019.

$$\mu_{thnx} = \mu_{xmas}$$

where μ_{thnx} and μ_{xmas} are the average Uber travel speeds for the selected Thanksgiving and Christmas days respectively in Seattle during 2018 and 2019.

Test Selection

As we had to compare the 2 festival groups, the unequal variance t-test also known as the Welch Test was an appropriate choice for testing the Null Hypothesis. The assumptions of this test listed below were also satisfied:

Check Performed	Status	Reference
Independence	Pass	3.4 Exploratory Data Analysis
Normality	Pass	4.2.2. Assumption Validation
Small sample size	Pass	4.2.2. Assumption Validation
Unequal variance	Pass	4.2.2. Assumption Validation

Table 6 - Assumptions and checks for test selection

Test Results

The results of the Welch test obtained in R are as follows:

- T-Statistic = -1.7572

- Df = 18.138
- P-Value = 0.09547
- Alternative hypothesis: true difference in means is not equal to 0
  95 percent confidence interval: (-3.1502962, 0.2797746)

Conclusion: Since p-value (0.095) is greater than our significance level (0.05), we do not have enough evidence to reject the Null Hypothesis of equal means for Thanksgiving and Christmas.

We also tested the Null Hypothesis using the equal variance t-test as the variances were not drastically different from each other, if not identical, (refer 4.2.2). We obtained a similar conclusion that the p-value (0.099) was greater than the significance level (0.05) and hence we did not have enough evidence to reject the Null Hypothesis.

4.2.4. Power and Sample Size

We were cognizant that conducting tests with very low sample sizes could result in low powers to detect the alternative hypothesis. Hence, we calculated the power to detect the difference between the true means for Thanksgiving (32.26) and Christmas (33.69) with an effect size of 1.43. This resulted in a very low power of 41.5% as per our initial guess. We further calculated the desired sample sizes to obtain 80% to 90% power and obtained the following:

Power	Desired Sample Size (per group)	Required years of data
80%	27	6
90%	36	7 to 8

Table 7 - Power and desired sample size

Since Uber only provides 2 years of data at present, this sample size per group was out of scope.

4.2.5. Inference

Caveat: Due to very low power to detect differences in means, as well as some deviation from the normality of data, we do not recommend making very strong conclusions based on the test results for the festival hypothesis.

Based on the results of the Welch Test, we cannot say with confidence that the mean travel speeds were different for Christmas and Thanksgiving periods in Seattle during 2018 and 2019. As per our descriptive statistics, we observed that the mean speeds were quite similar for these 2 festive periods.

Possible Reasons: Both Thanksgiving and Christmas festivals are known to be celebrated indoors with friends and family and people may only travel intermittently to reach their destinations. Hence there might be less traffic congestion on roads during both these festive periods due to which we observed the higher average speeds than usual in Seattle.

4.3. Peak Hours

We were interested in exploring how travel speeds may vary during the morning and evening peak hours which are influenced by office goers on weekdays. For this investigation, we chose to focus on morning and evening peak hours. We surveyed office goers in Seattle and identified 8 am to 11 am as the morning peak hours and 5 pm to 8pm as the evening peak hours.

4.3.1. Data Exploration

Trend Analysis

To build our understanding about the data, we plotted a trend chart (Refer Fig.13) for Uber's daily mean speeds over the morning peak hours and evening peak hours in 2019 to identify any underlying patterns or seasonality. Below are the key insights we took away:

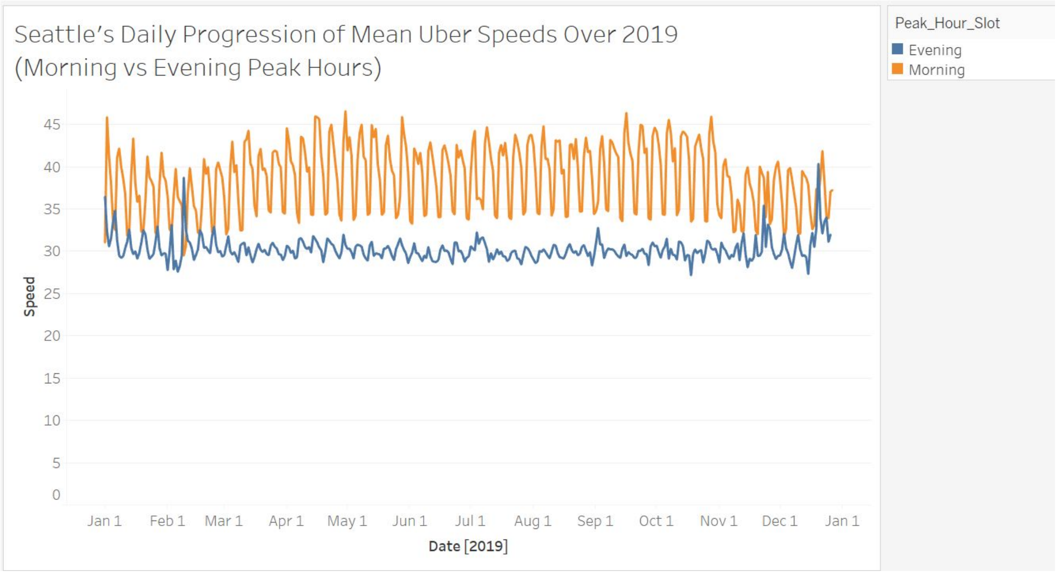


Fig.13 - Trends for Mean Uber speeds over morning and evening peak hours in 2019

Key Insights:

- The trend line for the morning peak hours shows a greater mean speed (40.68) as compared to the evening peak hours (30.05)
- The variance in the morning peak hour speeds (7.56) is much larger than in the evening peak hours (1.51)
- With a few exceptions, weekly seasonality can be observed in both morning and evening peak hours.

These insights were further validated by our descriptive statistics, boxplots and histograms.

The number of datapoints in both morning and evening peak hour groups is 771. Thus, the sample size can be said to be relatively large and is comparable. The data distribution of both morning and evening peak hours can be seen on the histograms below.

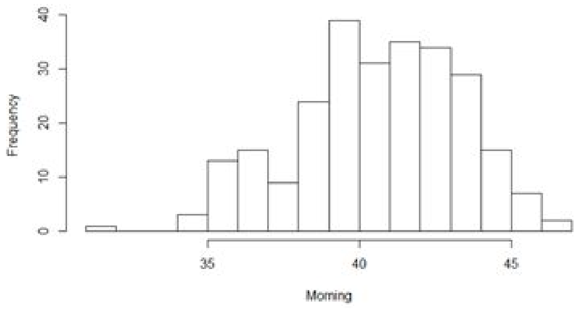


Fig.14 - Histogram for mean speeds for morning peak hours

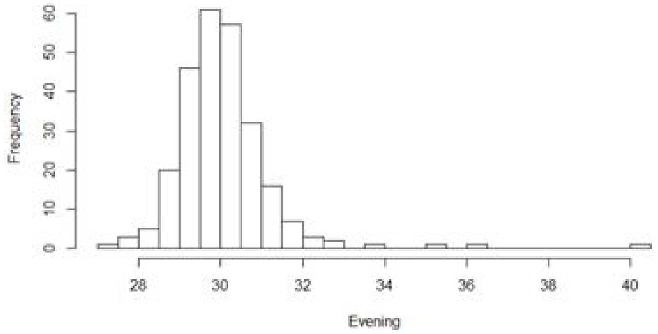


Fig. 15 - Histogram for mean speeds for evening peak hours

Statistic for **Mean Speed**	Morning	Evening
Median	40.90	29.91
Mean (μ)	40.68	30.05
Std. Deviation	2.74978	1.2287
Variance (σ^2)	7.56128	1.5098
Datapoints (n)	771	771

Table 8 - Descriptive statistics table speeds

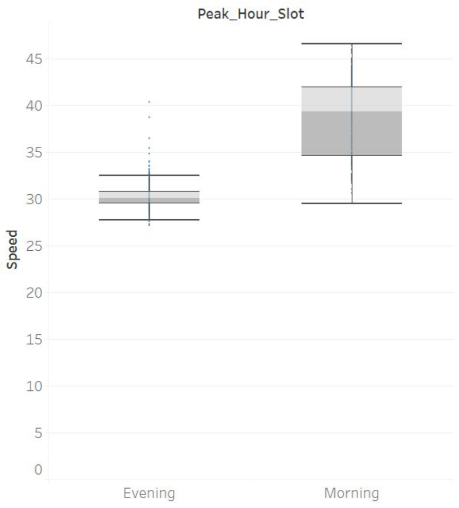


Fig 16. - Boxplot of morning and evening peak hour

Although both distributions are slightly skewed, due to large sample sizes, we were not required to check for the normality of the dataset.

Data Preparation

For this analysis, we had to compare data for speeds in the mornings and evenings on weekdays. To prepare data, we first filtered the data to just weekdays (Monday-Friday). Then, we filtered to just six hours of the day (8-11 am and 5-8 pm) and further aggregated the dataset on a daily level. Below is a snapshot of the prepared dataset.

Year	Month	Day	Day of Week	Hour	Peak Hour Slot	Mean Speed (mph)
2019	1	1	Tuesday	8	Morning	33.35249
2019	1	2	Tuesday	9	Morning	32.47098
...	...	...	...	...	...	...
2019	12	31	Tuesday	5	Evening	30.54495

Table 9 - Dataset snapshot for Hypothesis 3

4.3.2. Hypothesis Testing

Null Hypothesis

The average travel speed for Uber rides was the same on weekdays during morning and evening peak hours in Seattle in 2019.

$$\mu_{morning} = \mu_{evening}$$

where $\mu_{morning}$ and $\mu_{evening}$ are the average Uber travel speeds for morning and evening peak hours on weekdays respectively for Seattle in 2019.

Assumptions

Assumption / Check	Status	Reference
Independence	**Pass**	3.4 Exploratory Data Analysis
Normality/Large sample	**Pass**	4.3.1. Data Exploration

Table 10 - Assumptions and checks for test selection

Test Results

- Z-statistic = 60.00924
- Critical Z-Value Cutoff for 1-tail distribution at 0.05 significance level (α)= 1.64
- P-value < 2e-16
- Alternate Hypothesis: Mean morning peak hour speed is greater than mean evening peak hour speed
    - Power > 99% (Effect Size = 10.63, Sample Size = 771 per group)

Conclusion: P-value < 0.05, Reject the Null Hypothesis of equal means with 95% confidence. With a sample size of 771 per group, the test detects an effect size of 10.63 with over 99% power.

4.3.3. Inference

We can say with 95% confidence that the mean travel speeds are greater on weekdays for mornings as compared to evening peak hours in Seattle during 2019. This is also supported by our descriptive statistics.

We also tested this Null Hypothesis using the Welch test by simulating small samples of size 50 from the large sample and obtained the same conclusion of p-value < 0.05.

Possible Explanation: One possible reason for the difference in mean speeds could be because working professionals could be rushing to get to work early in the morning peak hours. This would explain the mean speeds in the morning being greater than in the evenings. Additionally, in the morning, there is likely only office traffic. But during evening peak hours, there could be many more reasons for congestion (eg: outings, social activity), which lead to lower speeds.

5. Discussion

5.1. Extended Inferences

The high level trends observed in Uber's movements could possibly be extrapolated to other automobile movements in Seattle. However, mean speeds might differ between Uber vehicles and other vehicles due to various factors like an enforced speed limit, Uber vehicles' car conditions and car types, as well as the representativeness of areas which Uber services. Hence we recommend that these observations and inferences should be restricted only to interpret Uber rides in Seattle.

Additionally, we cannot extend our analyses too much to the past or future years. This is because Uber is a relatively new company and forecasting the demand or speed restrictions for Uber rides is beyond the scope of our analysis. Further, we cannot gauge uniformity of the trend for traffic congestion. This is especially true in times of natural calamity (eg: coronavirus) where the traffic congestion has dropped significantly. Hence, it is ill advised to extrapolate past and future trends based on our analyses.

5.2. Limitations and Further Exploration

1. **Unavailability of vehicle counts on each street:** Due to this missing information, we could not take a weighted average while aggregating from street level data. We instead took a normal average, which is less representative of the underlying data.
2. **Mapping unavailable for OSM street IDs:** As we were unable to map street IDs in data to actual street names, we were unable to identify if a particular street was a highway or a local street and use that information.
3. **Missing data:** The data of movement trends in Seattle for 2 days of 2019 was missing.
4. **Excluded information to enforce privacy:** Uber purposely excluded information about those streets in Seattle which were travelled by less than 5 Uber vehicles in an hour to ensure privacy of the trips. Also, Uber clips off the starting segment as well as the ending segment of every ride while curating the movements dataset. This is done to preserve privacy of the origin place and the destination place of the Uber riders.
5. **Constrained data for Festival Hypothesis:** The data prepared for the analysis contains a total of 21 rows and is very small in size. Uber does not provide data prior to 2018 and hence we had to restrict our testing to this limited data sample. However, we decided to carry out the analysis as we wanted to take on the challenge of conducting the hypothesis testing in such real-life constrained settings where we may not have the luxury of having a large dataset.
6. **Alternative tests for small samples:** For the Festival Hypothesis, we tested the hypothesis using the Welch Test but obtained very low power due to small sample size. The test could be made more robust by applying other alternative methods such as Bootstrapping, Permutation t-test.

6. References

[1] University of Washington - Data 557 course slides by Prof. Brian

[2] User Movements Data Source
https://movement.uber.com/explore/seattle/speeds/query?dt[tpb]=ALL_DAY&dt[wd;]=1,2,3,4,5,6,7&dt[dr][sd]=2019-12-01&dt[dr][ed]=2019-12-31&ff=&lang=en-US

[3] Uber Movement: Speed Calculation Methodology
https://movement.uber.com/_static/56b3b1999eb80fadffbeb9bebe9888a7.pdf

7. Appendix

7.1. Code

Code for 'Day of Week' Hypothesis

```
#Outlier Treatment
library(tidyverse)
remove_outliers <- function(x) {
quantiles <- quantile(x, c(0, .95 ) )
y = x
y[ y > quantiles[2] ] <- quantiles[2]
y
}
add_new_column <- function(df) {
  speed_mph_mean_capped <-
remove_outliers(df$speed_mph_mean)
  return(cbind(speed_mph_mean_capped,df))
}
dfnew <- df %>%
  group_by(dayofweek) %>%
  nest() %>%
  mutate(data = map(data,
add_new_column)) %>%
  unnest()


#ANOVA
fit =
aov(speed_mph_mean_capped~factor(dayofweek),df)
summary(fit)


#Residual Diagnostics
residuals <- resid(fit)
boxplot(residuals ~ df$dayofweek, xlab =
"day of week")
```

Code for 'Festival' Hypothesis

```
 #Welch Test
 Thanksgiving <- all_thnx_data$mean_speed
 Christmas <- all_xmas_data$mean_speed
 t.test(Thanksgiving, Christmas, var.equal = FALSE)
```

```
#Equal Variance T-test
t.test(all_thnx_data$mean_speed,
all_xmas_data$mean_speed, var.equal = TRUE)
```

Code for 'Peak Hour' Hypothesis

```
#Z-test
s1=sd(morning$mean_speed)
s2=sd(evening$mean_speed)
n1=length(morning$mean_speed)
n2=length(evening$mean_speed)
Morning=morning$mean_speed
Evening=evening$mean_speed
X=mean(Morning)-mean(Evening)
mu0=0
se=sqrt(((s1*s1)/n1)+((s2*s2)/n2))
zstat=((X-mu0)/se)
data.frame(zstat,p=round((1-pnorm(zstat)),4))
```

```
#Code for Welch test (simulation)
set.seed(123456)
reps=20000
z=rep(NA,reps)
tstat=c(reps)
n1=50
n2=50
for(i in 1:reps)
{m=sample_n(morning, 50,
replace=TRUE) e=sample_n(evening, 50,
replace=TRUE)
tstat[i]=t.test(m,e,alternative="less",mu=0,
var.equal=F)$statistic}
welch.df = (s1^2/n1 + s2^2/n2)^2/
 (s1^4/(n1^2*(n1-1)) +
s2^4/(n2^2*(n2-1)))
mean(abs(tstat)>qt(.975,welch.df))
```

Power and Sample Size Calculations

```
#Power
alpha = 0.05    #significance level
n = nrow(total)/2    #sample size for each group
ma = mean(morning$mean_speed)
mb = mean(evening$mean_speed)
sa = sd(morning$mean_speed)
sb = sd(evening$mean_speed)
delta = ma - mb #difference between true means
t1 = -qnorm(1-(alpha/2))
t2 = abs(delta)/sqrt((sa*sa/n) + (sb*sb/n))
tf = t1 + t2
power = 100*pnorm(tf)
```

```
#Sample Size

# Desired Sample Size (80% Power)

beta = 1-0.80   #type 2 error probability
n <- ((sa^2 + sb^2) *
(qnorm(1-beta)+qnorm(1-alpha/2))^2)/
(delta^2)

n <- round(n,0)  #Desired sample size
```