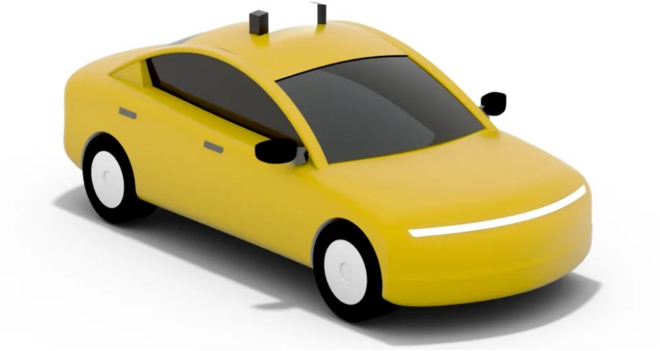
Uber Rides

Travel Speeds in Seattle

Authors: Aboli M, Harini R, Mayank G, Samarth M

Introduction

- Being frequent Uber riders in Seattle, we were interested in understanding the movement trends of Uber rides on the streets of Seattle.
- Uber Movement Speeds is a public dataset provided by Uber which contains anonymized travel data aggregated from their trips for major cities in the US for the year 2018 and 2019.
- We have focused on travel speed data of Seattle for the purpose of our analysis.



Data Profile

- **Size:** Over 46 million data points in one year, with a size over 15 GB. Due to the huge data size, the initial data procurement and processing was a very time intensive process in this project. We could appreciate the fact that to do any data science, almost 80% of the time needs to be invested in shaping the data correctly.
- **Granularity:** Hourly aggregated data of every Uber ride on the road segment with at least 5 unique trips in that hour.
- **Relevant Fields:** timestamp, year, month, day, and mean speed in mph.
- **Aggregation:** The speed aggregates have been taken by averaging the hourly speeds of Uber rides across different segments of streets in Seattle for the time periods (daily or hourly). This approximation has been chosen due to the lack of availability of other relevant traffic metrics like car counts per street, congestion at different times in a day, type of road (highway or not), type of car, etc.

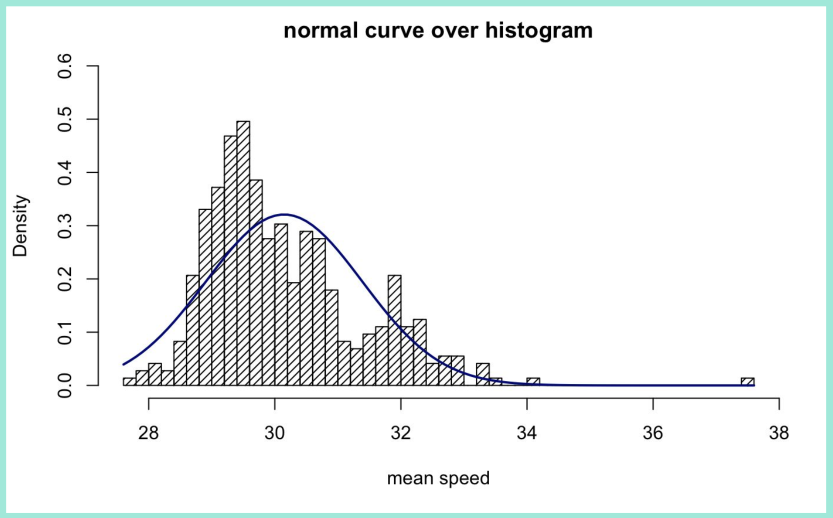
Caveats

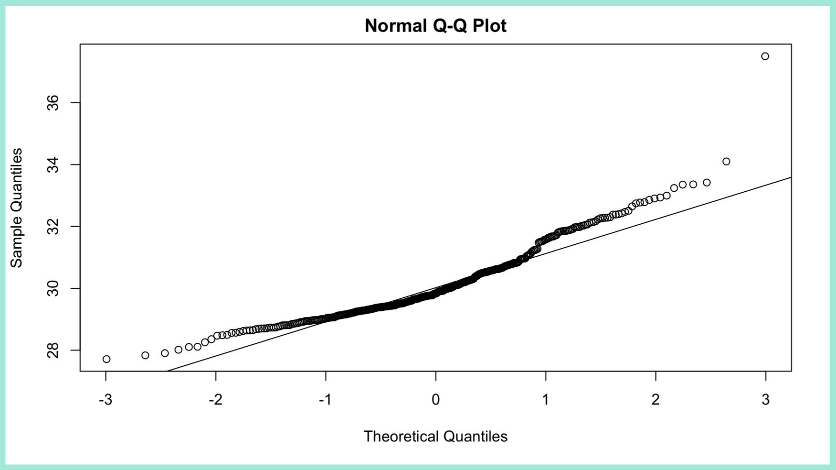
- Unavailability of vehicle count on each street to have a weighted average
- Unable to map the street IDs (Open Street Map IDs) to the corresponding streets in Seattle
- Few missing data points (about 2 days in 2019)
- Exclusion of streets travelled by less than 5 cars in an hour
- Clipping of the start and end segments of Uber car movements to preserve privacy of riders
- Based on Uber's data collection process, the data can be considered as a **randomized sample** of the entire population of Uber rides. Thus, the samples as well as the observations within the sample are **independent** of each other. Additional randomization can be accounted to the environmental factors, type of cars, traffic conditions, locality, etc.

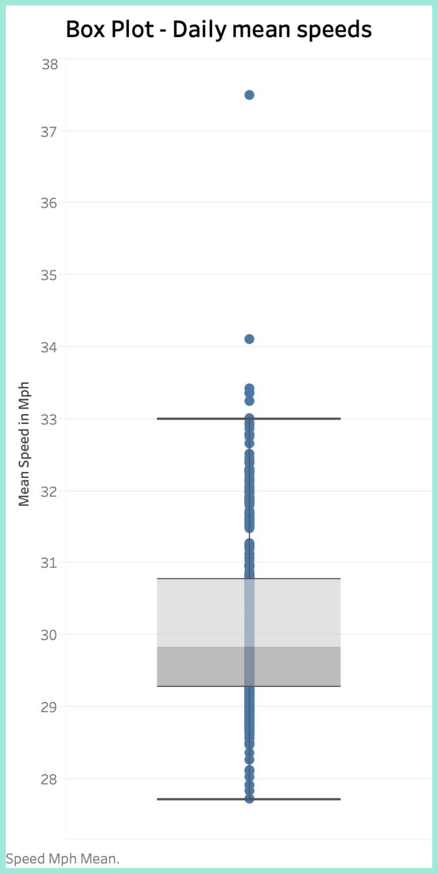
Exploratory Data Analysis

We observe an approximately **normal distribution** in the daily average speeds for Uber rides.

Not much **variance** is observed in the underlying data, with the exception of a few outliers.



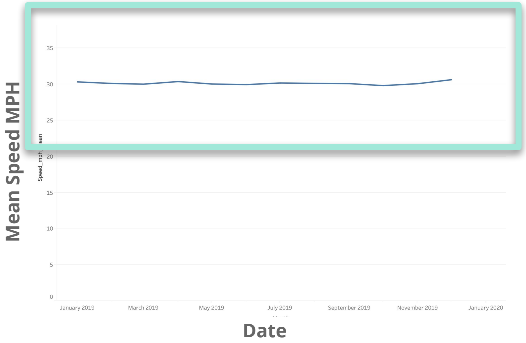




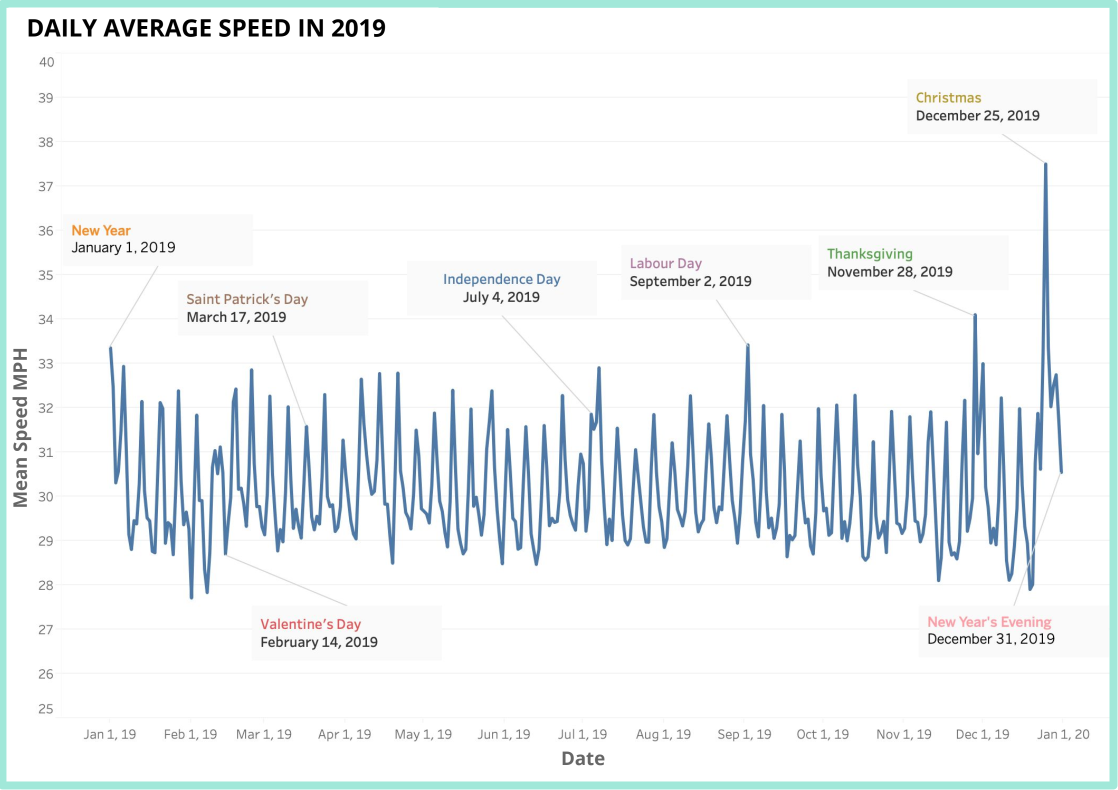
Trends

MEAN TRAVEL SPEED IN A YEAR

MONTHLY AVERAGE SPEED IN 2019



DAILY AVERAGE SPEED IN 2019



New Year
January 1, 2019

Saint Patrick's Day
March 17, 2019

Independence Day
July 4, 2019

Labour Day
September 2, 2019

Thanksgiving
November 28, 2019

Christmas
December 25, 2019

Valentine's Day
February 14, 2019

New Year's Evening
December 31, 2019

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

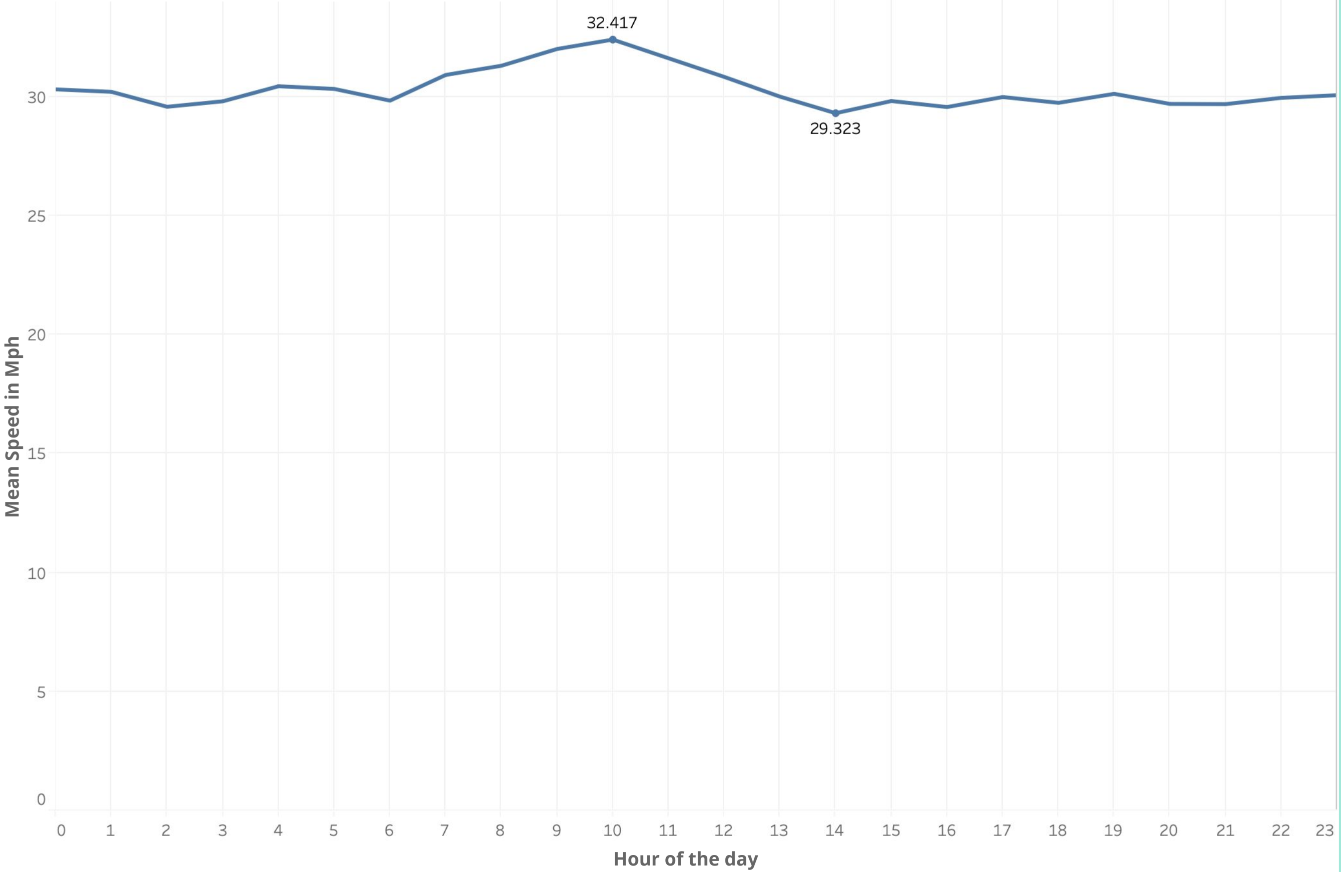
- **Seasonality** by the day of the week can be observed
- **Festive holidays** have some deviation from the average

MEAN TRAVEL SPEED IN A DAY

In 2019, **mornings** (6am to 12pm) tend to have higher average speed

From 2pm till midnight, the average speed remains approximately 30 mph

HOURLY AVERAGE SPEED IN 2019



Areas of Interest

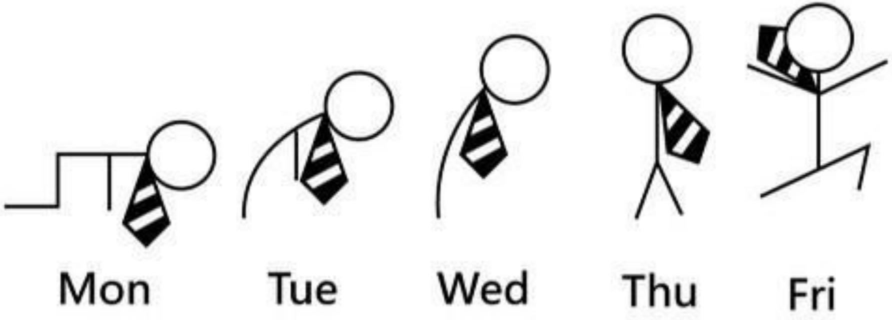
1. Seasonality by **day of week**
2. Impact of popular **festivals** on speeds
3. Speeds during **peak hours** of the day

Target Audience:

- Office goers
- People moving into Seattle
- Students
- Uber organisation
- Seattle city planning corporation
- Seattle police
- Anyone curious to understand the Uber movement trends during different times of the day, week and a year

Hypothesis 1

How does the average speed for Uber rides differ on each day of the week in Seattle in 2019?



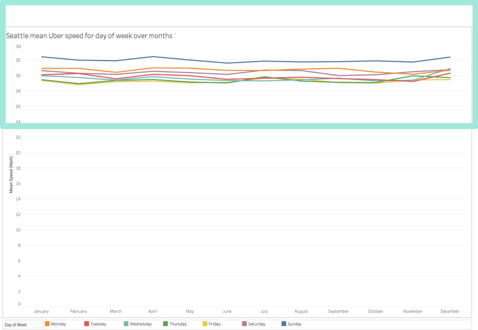
#MondayBlues

#WeekendVibes

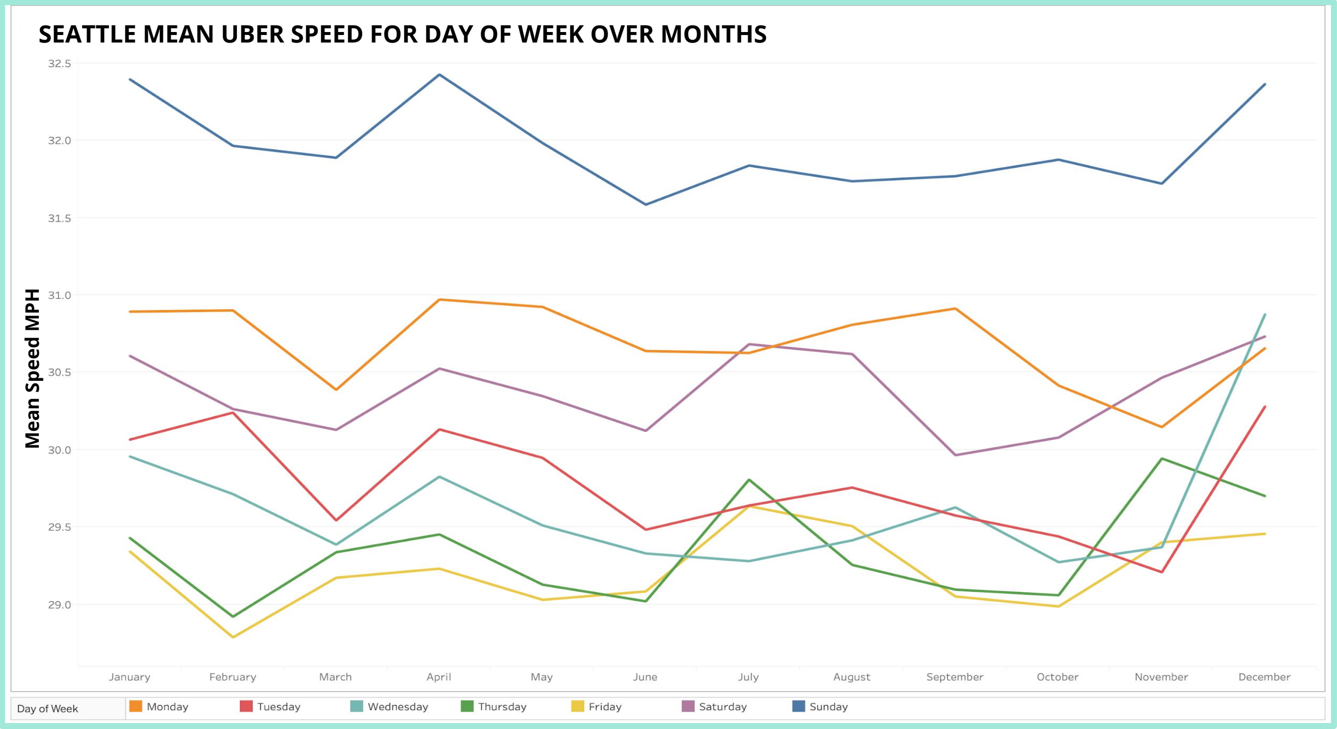
#ThankGodItsFriday

Trends

MEAN TRAVEL SPEED OVER MONTHS



Note: Minified view with Y-axis starting at 0



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

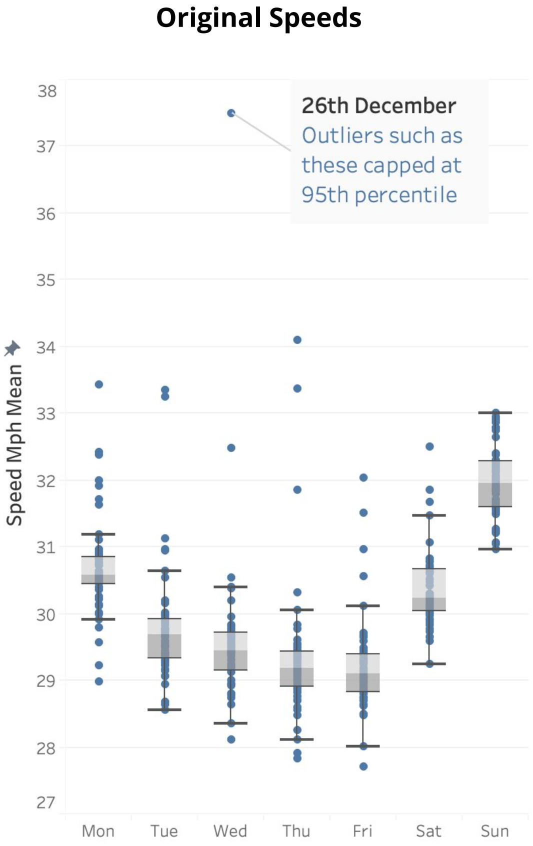
- **Sundays** are consistently faster around the year
- **Mondays** are surprisingly faster than Saturday and other weekdays

Variance

- We noticed some egregious **outliers**, resulting due to public holidays on weekdays and other miscellaneous factors.

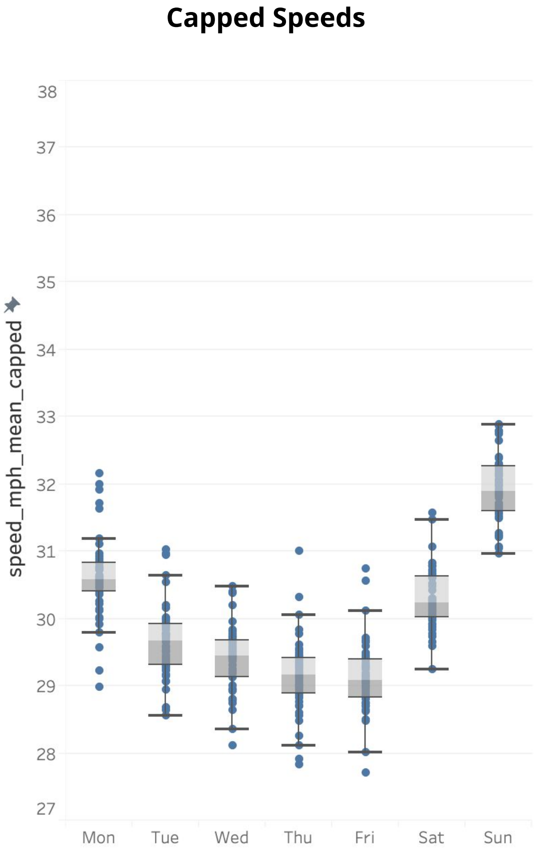
- For the purpose of our analysis, we decided to cap such outliers a the 95th percentile of that particular day.

- Variance becomes significantly more **consistent** after outlier treatment



Original Speeds

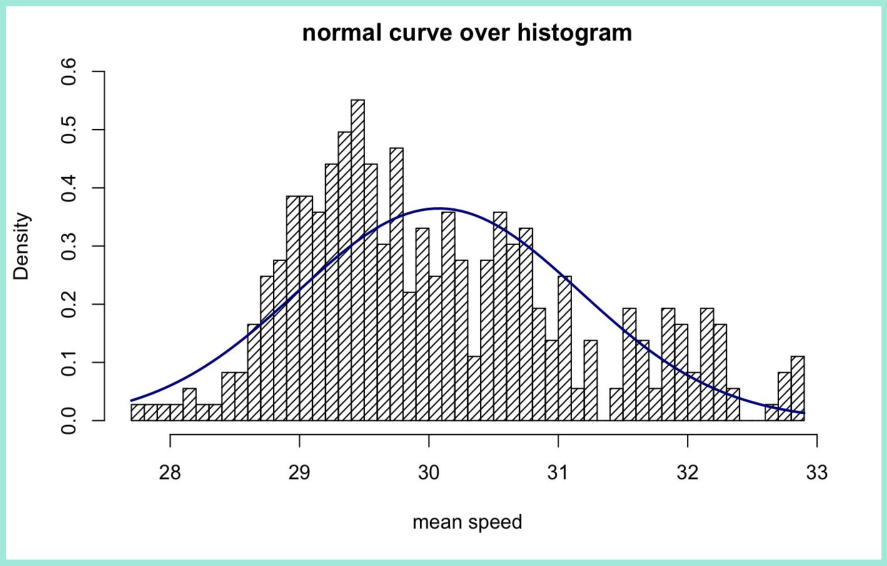
26th December
Outliers such as these capped at 95th percentile

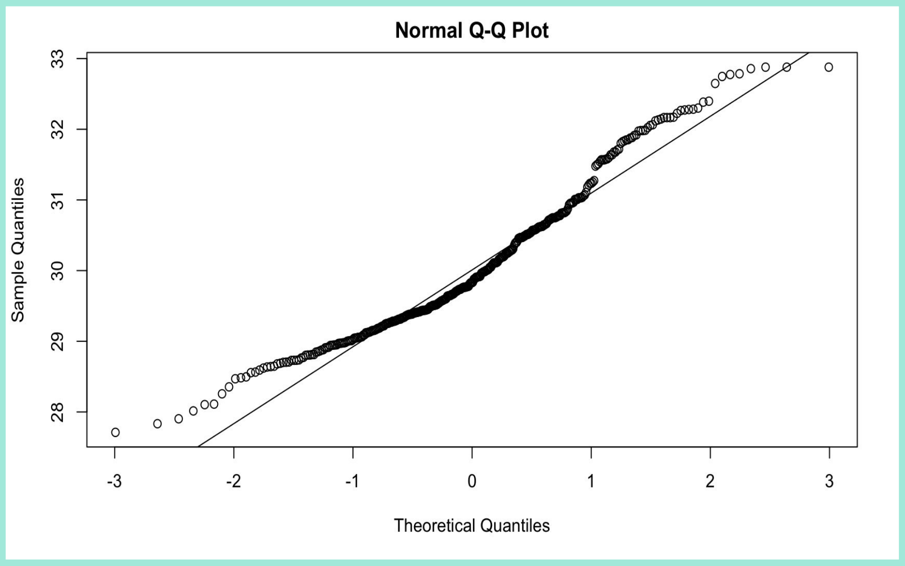


Capped Speeds

Normality

Histogram and Q-Q plot for population do not offer significant evidence against normal distribution assumption





Note: Both graphs are based on data points obtained after capping outliers at 95th percentile.

Null Hypothesis

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

$$\mu 1 = \mu 2 = \mu 3 = \mu 4 = \mu 5 = \mu 6 = \mu 7$$

where μ1 through μ7 are the average travel speeds for Monday through Sunday respectively for Seattle in 2019.

Hypothesis Testing

METHODOLOGY: Analysis of Variance (ANOVA)

ASSUMPTIONS

Independence ✓

Equal Variance ✓

Large Sample/Normality ✓

OUTCOME

F-STATISTIC = 152.8

P-VALUE = < 2e-16

```
                      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
```

RESULT: We can reject the null hypothesis at 95% confidence level

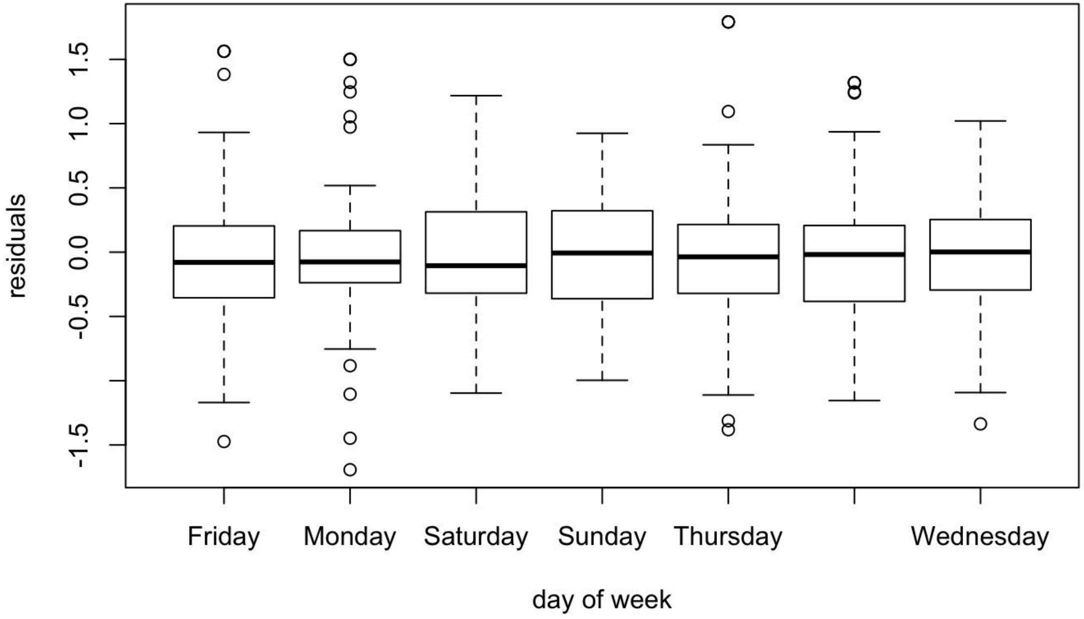
Note: R code used for this hypothesis testing can be found in Appendix 1.

Residual Diagnostic

RESIDUAL DIAGNOSTIC FOR EQUALITY OF VARIANCE

- Our equality of variance assumption was further corroborated by the boxplot showing variances of the residual for all 7 groups

- The plot can be interpreted as indicating **no evidence against equality of variances**.



Inferences

We can infer that **mean speed for Uber rides in Seattle in 2019 is not same for each day of the week**. This is in agreement with our results from descriptive statistics and trends analysis. Important to note that **mean speed is taken across streets without accounting for number of cars**.

To drill down further, we also performed 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.

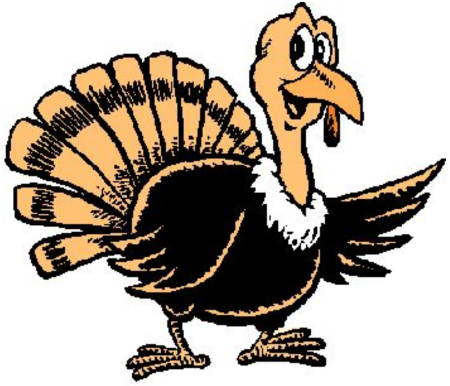
Possible Explanation: The result is intuitive and seems logical as **traffic is highly influenced by work commute**, which affects the mean speed on working days, but not on weekends. It can further be influenced by factors such as employees working from home, different office timings for Fridays, people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

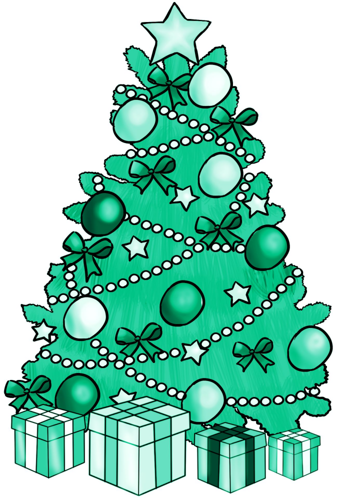
Hypothesis 2

Is the average travel speed for Uber rides comparable during the different festive times in Seattle?

We would like to compare the 2 most popular holidays in Seattle for this analysis:

Thanksgiving vs. **Christmas**

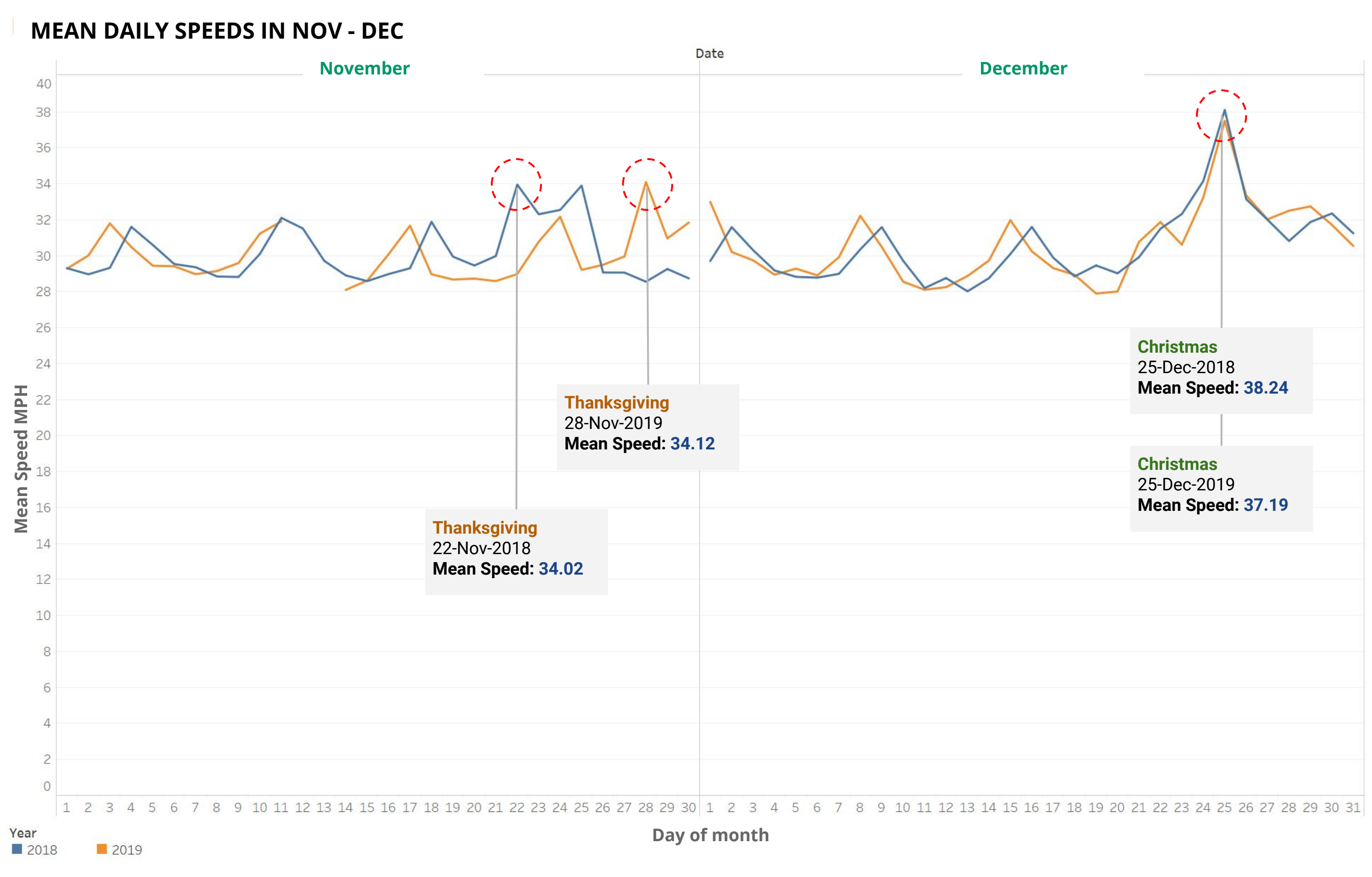




Trends

Mean speed is **higher** on days of Thanksgiving and Christmas compared to other days.

This observation made us **curious** to test if there is a **significant difference** between speeds during the 2 **festive holiday** periods.


MEAN DAILY SPEEDS IN NOV - DEC

Thanksgiving
22-Nov-2018
Mean Speed: 34.02

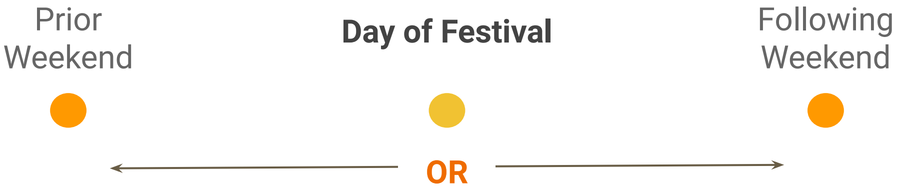
Thanksgiving
28-Nov-2019
Mean Speed: 34.12

Christmas
25-Dec-2018
Mean Speed: 38.24

Christmas
25-Dec-2019
Mean Speed: 37.19

Dataset

FESTIVE PERIOD



Thanksgiving: **10 Days**

In **2018**: Wed, 21 Nov to Sun, 25 Nov

In **2019**: Wed, 27 Nov to Sun, 01 Dec

Christmas: **11 Days**

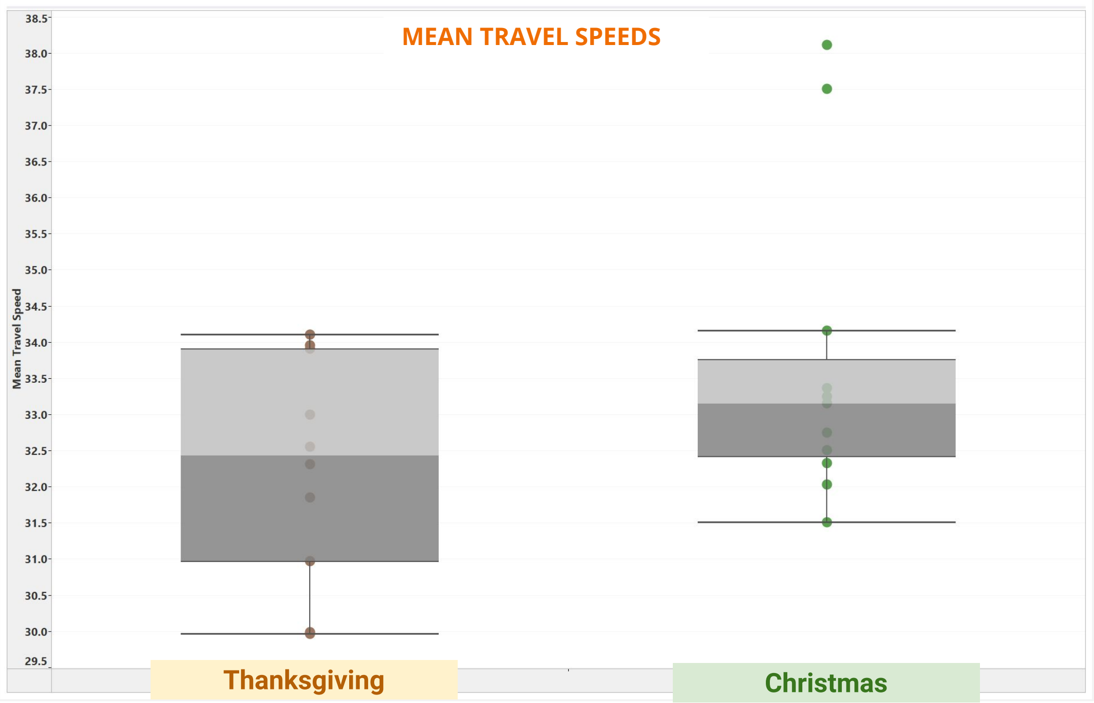
In **2018**: Sat, 22 Dec to Wed, 26 Dec

In **2019**: Tue, 24 Dec to Sun, 29 Dec

- The **sample size** for both groups in this case is **very small** and **comparable** to each other.
- We wanted one of our hypotheses to be on a small size dataset to understand how testing can be applied in a **constrained real-life setting.**

Descriptive Statistics

Mean Speed	Thanksgiving	Christmas
Min.	29.96	31.51
1st Qu.	31.19	32.41
Median	32.43	33.15
Mean	**32.26**	**33.69**
3rd Qu.	33.68	33.76
Max.	34.1	38.11
Std Deviation	1.56	2.16
Variance	**2.43**	**4.66**
Datapoints	**10**	**11**



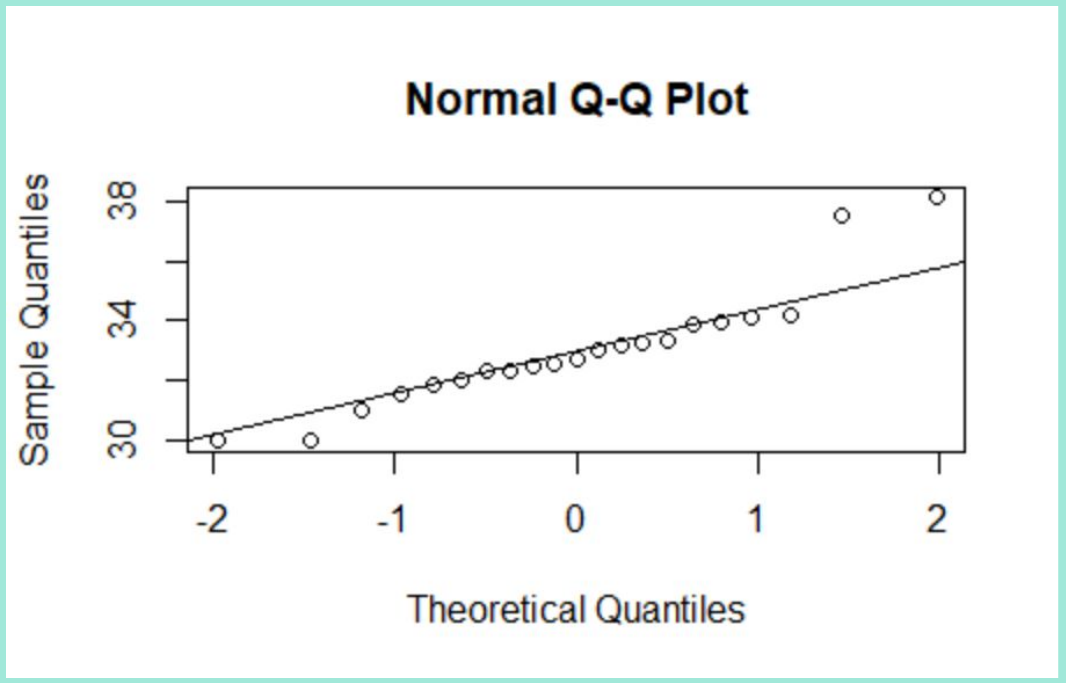
- The **variances** of the 2 groups are **unequal** but the **mean** speeds are quite **similar**.
- **Christmas** period shows **higher variance** in mean speeds as compared to Thanksgiving period.

Normality

Earlier, we have earlier observed that the **population** follows a **normal** distribution.

Here we can observe that the **sampled data** for the festivals also follows a **normal** distribution with the exception of a few outliers.

Refer appendix for R code to generate this plot.



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.

$$\mu 1 = \mu 2$$

where μ1 and μ2 are the average travel speeds for the selected Thanksgiving days and Christmas days in Seattle for 2018 and 2019.

Hypothesis Testing

METHODOLOGY: Welch Test (Unequal variance t-test)

Refer Appendix for R Code

CHECKS

Independence	✓
Unequal Variance	✓
Normality	✓
Small Sample Sizes	✓

OUTCOME

T-Statistic = -1.7572, Df = 18.138, P-Value = 0.09547

Power = 41.5% (Effect size = 1.43)

Desired Sample Size for 80% Power = 27 per group

Data: Thanksgiving and Christmas
Alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: (-3.1502962 0.2797746)
Sample of estimates:
Mean of x: 32.25829, Mean of y: 33.69355

RESULT: P-value > 0.05, not enough evidence to reject the Null Hypothesis of equal means

Hypothesis Testing

METHODOLOGY: Welch Test (Unequal variance t-test)

Refer Appendix for R Code

Very low power (41.5%) due to a **small sample size** (10 per group)

Desired sample sizes:
80% power → 27 samples per group
90% power → 36 samples per group

Uber does not provide this much data and hence this result should be interpreted with more caution.

CHECKS

Independence	✓
Unequal Variance	✓
Normality	✓
Small Sample Sizes	✓

OUTCOME

T-Statistic = -1.7572, Df = 18.138, P-Value = 0.09547

Power = 41.5% (Effect size = 1.43)

Desired Sample Size for 80% Power = 27 per group

Data: Thanksgiving and Christmas
Alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: (-3.1502962 0.2797746)
Sample of estimates:
Mean of x: 32.25829, Mean of y: 33.69355

RESULT: **P-value > 0.05, not enough evidence to reject the Null Hypothesis of equal means**

Inference

Caveat: Due to **very low power** to detect differences in means as well as **some deviation from normality**, we should not draw very strong conclusions based on this test result.

As per our test results, we **cannot** say with confidence that the **mean travel speeds** were **different** for Christmas and Thanksgiving 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.

We also tested this Null Hypothesis using the **equal variance t-test** as the variances were not drastically different from each other (if not identical) and obtained **similar results (p-value > 0.05)**

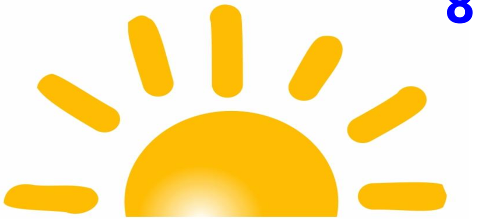
Possible Reasons: These festivals are known to be celebrated **indoors** and people may **travel intermittently** to reach their destinations.

Hypothesis 3

How did travel speed differ for Uber rides on weekdays during the peak hours of the morning vs. peak hours of the evening in Seattle in the year 2019?

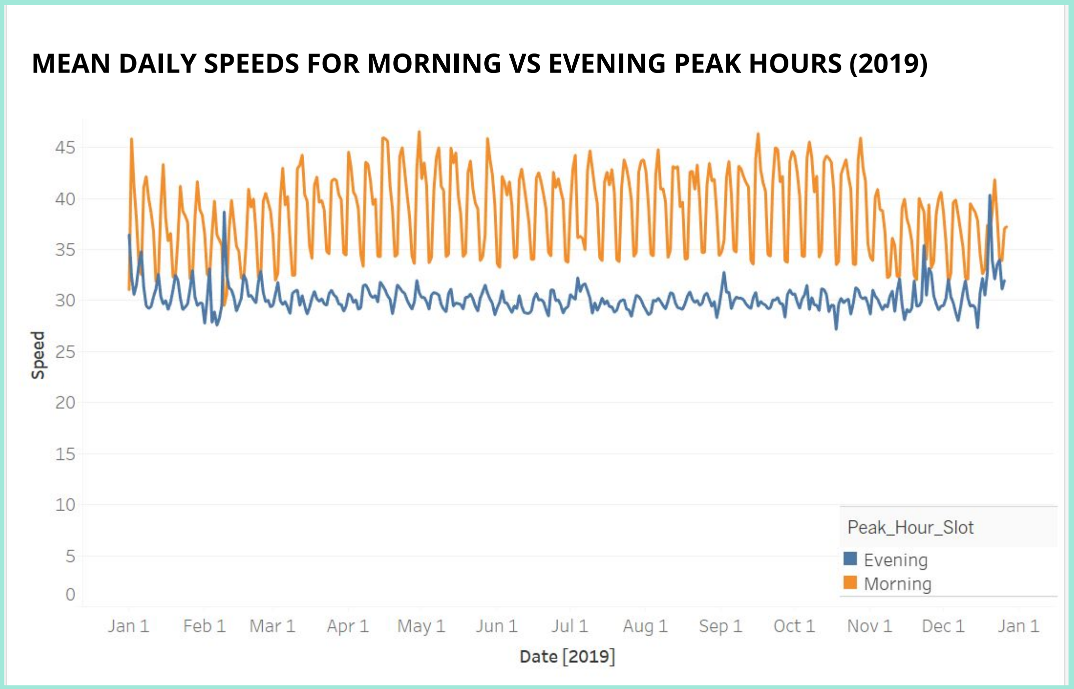
Morning Peak Hours vs. **Evening Peak Hours**

8 am to 11 am vs. **5 pm to 8 pm**



Trends

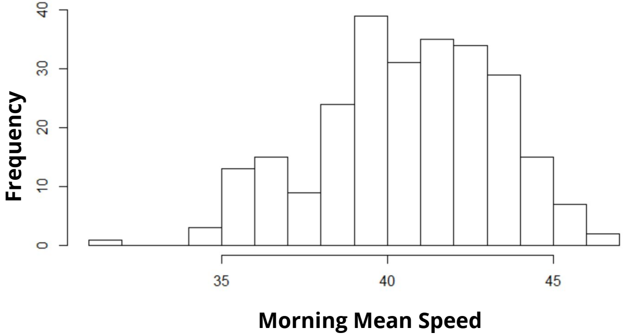
Summary	Morning	Evening
Min.	31.02	27.30
Mean	**40.68**	**30.05**
Max.	46.57	40.33
Std Deviation	2.75	1.22
Variance	**7.56**	**1.51**
Datapoints	**771**	**771**



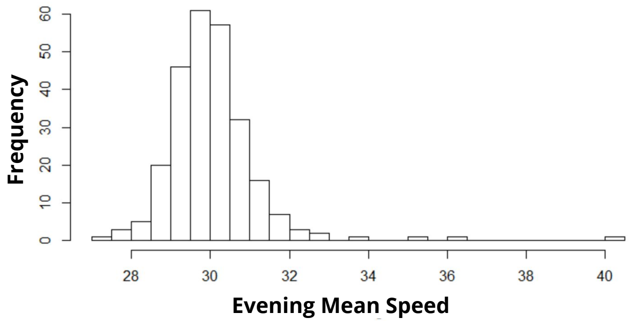
- Much **larger variance** in **morning peak hour speeds** than in evening peak hours
- **Mean speeds:** Morning > Evening
- Sample size is relatively **large** and **comparable**

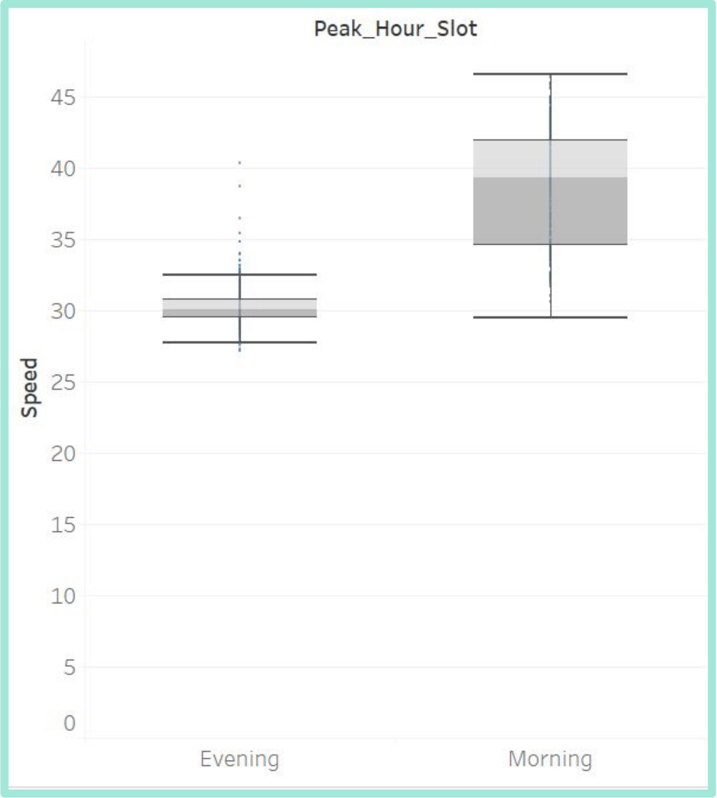
Data Distribution

MEAN SPEED - MORNING PEAK HOURS



MEAN SPEED - EVENING PEAK HOURS





The **morning distribution** looks slightly **right skewed with greater variance** whereas the **evening distribution** looks slightly **left skewed**

Null Hypothesis

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

$$\mu 1 = \mu 2$$

where μ1 and μ2 are the average travel speeds for morning and evening peak hours for weekdays in Seattle for 2019.

Hypothesis Testing

METHODOLOGY: Large sample z-test (one-sided)

Refer Appendix for R Code

ASSUMPTIONS

Independence ✓

Large Sample/Normality ✓

OUTCOME

Z-STATISTIC = 60.00924

Compare against: Critical Z-Value 1.64 for 1-tail distribution at 0.05 significance level

P-VALUE < 2e-16

POWER > 99% **(Effect Size = 10.63, Sample Size = 771 per group)**

RESULT: P-value < 0.05, Reject the Null Hypothesis of equal means with 95% confidence.

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** where we examined that the mean speeds for the morning peak hours was greater than evening peak hours.

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. 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). This could be another reason as to why evening peak hours' mean speeds may be lower.

Extended Inferences

- The high level trends observed in Uber's movements could possibly be extrapolated to other automobile movements in Seattle.
- However, the mean speeds between Uber vehicles and other vehicles in Seattle might differ due to various factors such as:
    - Uber drivers need to follow a speed limit and driving rules
    - Car condition of Uber vehicles is monitored
    - Car type of Uber vehicles is usually Sedans, Hatchbacks and SUVs. This is not representative of other vehicles such as pickup trucks, buses, motorbikes in Seattle.
    - Localities in which Uber might be providing services is not representative of the entire Seattle area
- **Hence we recommend that these observations and inferences should be restricted only to interpret Uber rides in Seattle.**

THANK YOU!

QUESTIONS?

Appendix

Appendix: R Codes for Analysis

Hypothesis 1 (Days of Week)

Code for 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()
```

Code for ANOVA test

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

Code for residual check

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

Hypothesis 2 (Festivals)

Code for Welch Test

```
Thanksgiving <- all_thnx_data$mean_speed
Christmas <- all_xmas_data$mean_speed
t.test(Thanksgiving, Christmas, var.equal = FALSE)
qqplot(y=all_xmas_data$mean_speed)
```

Code for ANOVA

```
summary(aov(mean_speed ~ factor(flag), data=all_festival_data))
```

Code for Equal Variance T-test

```
t.test(all_thnx_data$mean_speed, all_xmas_data$mean_speed, var.equal = TRUE)
```

Hypothesis 3 (Peak Hours)

Code for 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)
{
  morning_sample=sample_n(morning,50,replace=TRUE)
  evening_sample=sample_n(evening,50,replace=TRUE)
tstat[i]=t.test(morning_sample,evening_sample,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 Calculations and Sample Size

Power Calculation

```
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)           #power
print(power)
```

Sample Size Calculation

```
# 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)
print(n)                               #Desired sample size

# Desired Sample Size (90% Power)
beta = 1-0.90                           #type 2 error probability
n <- ((sa^2 + sb^2)*(qnorm(1-beta)+qnorm(1-alpha/2))^2)/(delta^2)
n <- round(n,0)
print(n)                               #Desired sample size
```

Descriptive Statistics

Histogram

~~~
g = df$speed_mph_mean
m<-mean(g)
std<-sqrt(var(g))
hist(g, density=20, breaks=40, prob=TRUE,
    xlab="mean speed", ylim=c(0, 0.6),
    main="normal curve over histogram")
curve(dnorm(x, mean=m, sd=std),
     col="darkblue", lwd=2, add=TRUE, yaxt="n")
```
~~~

Box Plot

```
boxplot(df$speed_mph_mean_capped~df$dayofweek)
```

QQ Plot

```
qqnorm(df$speed_mph_mean)
qqline(df$speed_mph_mean)
```

Appendix: Data Details

Day of Week Dataset

Snapshot of the aggregated dataset for this hypothesis

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

Data Summary (Mean Speed)

Mean	30.15
SD	1.24
Var	1.54
Min	27.71
Max	37.50

Note: These averages have been taken by averaging the hourly speeds of Uber rides across different segments of streets in Seattle for the selected dates. This approximation has been chosen due to the lack of availability of other relevant traffic metrics like car counts per street, congestion at different times in a day, type of road (highway or not), type of car, etc.

Day of Week Descriptive Statistics

Notice the **change in variances between uncapped and capped** data points. As variances for capped data are much closer to each other and are also relatively low compared to the mean, we felt confident in assuming **equality of variance**.

Day of week	Mean	Var (uncapped)	Var (capped)	Min	Max	Sample size
Monday	30.66488	0.5808846	0.4467365	28.97204	32.16502	52
Tuesday	29.71113	0.7862165	0.3533132	28.55620	31.02958	52
Wednesday	29.44743	1.6810621	0.2717846	28.11119	30.46823	51
Thursday	29.21528	1.1641402	0.4258321	27.83370	31.00691	52
Friday	29.18465	0.5453930	0.3635415	27.71119	30.74745	52
Saturday	30.34955	0.3365615	0.2547196	29.25290	31.56715	52
Sunday	31.95395	0.2739123	0.2662607	30.95696	32.87888	52

Festival Dataset Snapshot

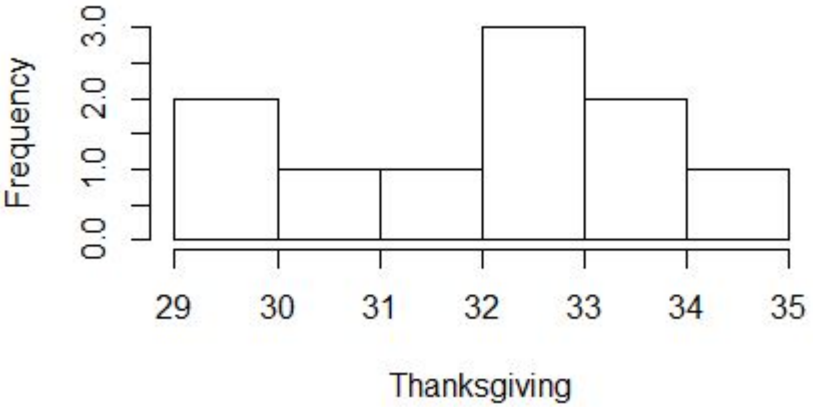
A snapshot of the filtered and aggregated dataset prepared for this hypothesis

Year	Month	Day	Festival	Mean Speed (mph)
2018	11	21	Thanksgiving	29.98616
2018	11	22	Thanksgiving	33.95708
2018	11	23	Thanksgiving	32.31070
...	...	...	...	...
2019	12	29	Christmas	32.74694

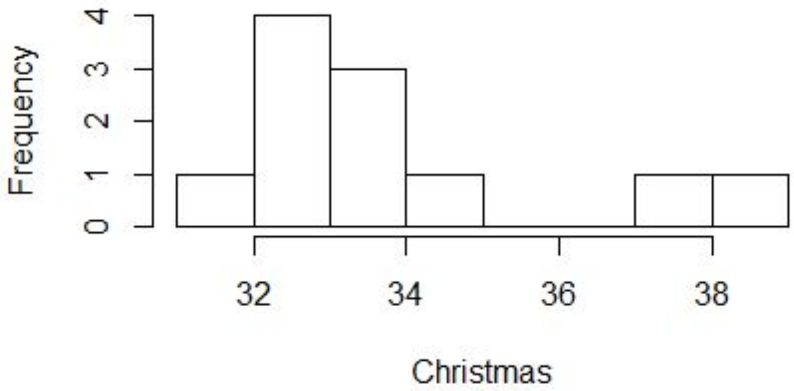
Note: The speed aggregates have been taken by averaging the hourly speeds of Uber rides across different segments of streets in Seattle for the selected dates. This approximation has been chosen due to the lack of availability of other relevant traffic metrics like car counts per street, congestion at different times in a day, type of road (highway or not), type of car, etc.

Festival Data Distributions

MEAN SPEED DURING THANKSGIVING



MEAN SPEED DURING CHRISTMAS



- Due to very small sample sizes, it is difficult to determine the **distribution** for the 2 groups.
- However, the distribution of the underlying population is **normal** (Refer)

Peak Hours Dataset

Morning: 257 Days

3 Hours: 8 am -10 am

Sample size: 771

Evening: 257 Days

3 Hours: 5 pm - 8 pm

Sample size: 771

- Data missing for certain days / peak hours (**28** missing data points)
- The sample sizes in both cases are **large** and **comparable** to each other

Peak Hours Dataset Snapshot

Snapshot of the aggregated dataset for this hypothesis

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	1	3	Tuesday	10	Morning	30.30966
...	...	...	...		...	...
2019	12	31	Tuesday		Evening	30.54495

Note: These averages have been taken by averaging the hourly speeds of Uber rides across different segments of streets in Seattle in 2019, for the selected six peak hours for weekdays. This approximation has been chosen due to the lack of availability of other relevant traffic metrics like car counts per street, congestion at different times in a day, type of road (highway or not), type of car, etc.