DATA 557: FINAL PROJECT PROPOSAL WITH ANALYSIS PLAN

GROUP DETAILS

- Aboli Moroney
- Harini Ramprasad
- Mayank Goel
- Samarth Modi

PRELIMINARY HYPOTHESIS

Some of the potential questions which we seek to find answers to are as follows:

- How did the travel speed differ on each day of the week in Seattle in 2019?
- Is the travel time comparable during the different festive days in Seattle? We would like to consider holidays for Christmas vs. Thanksgiving in 2018 and 2019 for this analysis.
- How did travel speed differ on weekdays during the peak hours of the morning (8 am to 11am) vs. peak hours of the evening (5 pm to 8 pm) in Seattle in the years 2018 and 2019?

DATA DETAILS

I. Uber Speed Data

This dataset provides the average speed and travel-time on a given road segment for each hour of each day in the specified period. It only includes road segments with at least 5 unique trips in that hour. Dataset is available for download at the following link:
https://movement.uber.com/cities/seattle/downloads/speeds?lang=en-US&tp[y]=2019&tp[q]=4

II. Data Quality Analysis

- A reasonably large dataset with millions of data points at a daily level from 2018 to 2019 worldwide for all countries where Uber is providing services, maintained by Uber
- No cleaning required, no missing values

III. Data Schema

Column Name	Description	Type
year	Year (local city time)	int
quarter	Quarter 1-4 (local city time)	int
hour_of_day	Hour of day 0-23 (local city time)	int
osm_way_id	Corresponding OpenStreetMap Way ID for this segment. Note that one OpenStreetMap Way may contain multiple Movement segments.	bigint
osm_start_node_id	Corresponding OpenStreetMap Node ID for this junction.	bigint
osm_end_node_id	Corresponding OpenStreetMap Node ID for this junction.	bigint
speed_mph_mean	The average speed of Uber vehicles on this road segment in mph	float
speed_mph_stdev	The standard deviation of speeds on this road segment in mph	float
speed_mph_p50	50th percentile (median) speed of Uber vehicles on this road segment in mph	float

speed_mph_p85	85th percentile speed of Uber vehicles on this road segment in mph	float

FEEDBACK AND REVISIONS

Based on the feedback from TAs and Prof.Brian, we have further refined our hypothesis and approach as follows:

- We will do some preliminary exploration of the data to identify the outliers and treat them before running hypothesis tests to ensure that the groups under consideration are comparable.
- We have decided to drop the hypothesis for comparison travel times in Seattle vs San Francisco vs New York as we do not have an accurate way to aggregate street-level data to obtain city-level average speeds.
- We have decided to add an additional hypothesis regarding the travel speeds in peak hours in the morning versus the evening, as we found this to be an interesting question to answer. This will be identified as hypothesis 2.

ANALYSIS PLAN

Hypothesis I

How did the travel speed differ on each day of the week in Seattle in 2019?

Null Hypothesis: The average travel speed for Uber rides was the same for all days of the week in Seattle in 2019

Ho: m1= m2 = m3 = m4 = m5 = m6 = m7; where m1 through m7 are the average travel speeds from Monday through Sunday.

Assumptions

- Randomization: We are assuming Uber's average speeds are similar to the rest of the city car traffic. Hence, Uber speed data can be considered as a random sample of the entire population for Seattle city car speed data.
- Equal variance: Assuming equal variance, to be updated based on analysis
- Normality: We will test the normality of underlying data using ggplot and histograms. However, since we have a reasonably large number of samples for each day, this assumption should not be required.

Data

- The downloaded dataset from the source contains daily average speed at street level in Seattle for each day of the week for 2019.
  - Columns: year, month, day, day_of_week, hour, speed_mph_mean, osm_start_node_id, osm_end_node_id
  - Uniqueness: year, month, day, osm_start_node_id, osm_end_node_id
- To prepare the dataset for analysis, we will aggregate the speeds to get the mean speed for Seattle for each date using 'average' as an aggregation. Below is a snapshot of the prepared dataset which will be unique on the 'date' field.

date	day_of_week	speed_mph_mean
01-Jan-2019	2	60
02-Jan-2019	3	65

03-Jan-2019	4	80
...	...	...
31-Dec-2019	2	75

- Since we are taking the data only for 2019, we expect to have a total of 365 data points with approximately 52 data points for each day of the week.

Analysis

Based on the validation of assumptions, we would choose an ideal hypothesis test. If we are able to assume the equal variance, we will perform the ANOVA test where each day of the week will be a level in the factor.

Hypothesis II

How did travel speed differ on weekdays during the peak hours of the morning (8 am to 11am) vs. peak hours of the evening (5 pm to 8 pm) in Seattle in the years 2018 and 2019?

Null Hypothesis: The average travel speed for Uber is the same on weekdays for both morning and evening peak hours, across 2018 and 2019.
Ho: s1= s2; where s1 and s2 are the average travel speeds for morning peak hours (8-11 am) and evening peak hours (5-8 pm) across 2018 and 2019.

Assumptions

- Randomization: We are assuming Uber's average speeds are similar to the rest of the city traffic. Hence, Uber speed data can be considered as a random sample of the entire population for Seattle city speed data
- Equal variance: Assuming equal variance, to be updated based on analysis
- Normality: We will test the normality of underlying data using gg-plot and histograms. However, since we have a reasonably large number of samples for each day, this assumption should not be required.
- The peak hours in Seattle during the morning are from 8-11 am and in the evening are from 5-8 pm. We are assuming this based on our experiences living in Seattle and conversations with multiple office going people in Seattle.

Data

- The downloaded dataset from the source contains hourly average speed at street level in Seattle for each day in 2018 and 2019.
    - Columns: year, month, day, hour, speed_mph_mean, osm_start_node_id, osm_end_node_id, day_of_week
    - Uniqueness: year, month, day, hour, osm_start_node_id, osm_end_node_id
- To prepare the dataset for analysis, we will download data filtered for all weekdays of 2018 and 2019. Additionally, we will filter the data for the morning and evening peak hours and label these under 'peak_hour_slot'. We will take an aggregate average of speeds in both morning and evening hours for our analysis. Below is a snapshot of the prepared dataset which will be unique on the combination of 'date' and 'peak_hour_slot' fields.

date	**day_of_week**	**peak_hour_slot**	**speed_mph_mean**
01-Jan-2019	2	morning	60

01-Jan-2019	2	morning	65
01-Jan-2019	2	morning	80
01-Jan-2019	2	evening	70
01-Jan-2019	2	evening	55
...	...	...	...

- Thus we would get approximately 3120 samples (6 hours * 5 days * 52 weeks * 2 years) for comparison.

Analysis

Based on the validation of assumptions, we would choose an ideal hypothesis test. If we are able to assume the equal variance, we will perform the equal variance t-test for the field ‘peak_hour_slot’ as factor and ‘evening’ and ‘morning’ as factor levels.

Hypothesis III

Is the travel time comparable during the different festival days in Seattle? We would like to consider holiday days of Christmas vs Thanksgiving in 2018 and 2019 for this analysis.

Null Hypothesis: The average travel speed for Uber was the same for both festive days of Christmas and Thanksgiving in Seattle across 2018 and 2019.
Ho: f1= f2; where f1 and f2 are the average travel speeds from Thanksgiving days and Christmas days

Assumptions

- Randomization: We are assuming Uber’s average speeds are similar to the rest of the city traffic. Hence, Uber speed data can be considered as a random sample of the entire population for Seattle city speed data
- Equal variance: Assuming equal variance, to be updated based on analysis
- Normality: We will test the normality of underlying data using gg-plot and histograms. However, since we have a reasonably large number of samples for each day, this assumption should not be required.

Data

- The downloaded dataset from the source contains hourly average speed at street level in Seattle for each day in 2018 and 2019.
    - Columns: year, month, day, hour, speed_mph_mean, osm_start_node_id, osm_end_node_id
    - Uniqueness: year, month, day, hour, osm_start_node_id, osm_end_node_id
- We will consider the following days as holiday dates for the 2 festivals:
    - Christmas: 24 to 26 Dec, 2018 and 24 to 26 Dec, 2019 (6 days across both years)
    - Thanksgiving: 21 to 23 Nov, 2018 and 27 to 29 Nov, 2019 (6 days across both years)
    - We have thoughtfully considered 1 additional day before and after the actual festivals as we want to consider the effect around the festive time which is often more than just the actual day.
- To prepare the dataset for analysis, we will filter the data for the above dates and use it at an hourly level for our analysis. Below is a snapshot of the prepared dataset which will be unique on the ‘date’ field.

festival	date	hour_of_day	speed_mph_mean
Christmas	24-Dec-2018	12:00	70
Christmas	24-Dec-2018	01:00	65
...	...	...	80
Thanksgiving	29-Nov-2019	11:00	50
Thanksgiving	29-Nov-2019	12:00	75

- Thus we would get approximately 144 samples (24 hours * 6 days) for both the festivals for comparison.

Analysis

Based on the validation of assumptions, we would choose an ideal hypothesis test. If we are able to assume the equal variance, we will perform the ANOVA test where each Festival would be a level in the factor.