Northeastern University
College *of* Engineering

Engineering Probability and Statistics
IE 6200
Section 06
December 2, 2019

Effect of Weather Conditions on Uber pickups in New York City

Priyanka Tadse

Varun Pant

Hanoz Patel

Table of Contents

1. Problem Statement

Travel is a part and parcel of the life of every human being. One may travel places for work, school, meetings or just to visit a place. There are various means of transportation to aid this travel like personal cars, trains, buses, flights or cab services. One such private cab service is Uber. In this project, we have explored the effects of weather on the number of Uber pickups in 5 boroughs of New York City.

2. Data Collection

The dataset we will be using for this project includes the data of Uber rides in and around New York from the month of January to June 2015 obtained from Kaggle under the title 'NYC Uber Pickups with Weather and Holidays'. It contains the number of pickups for every single hour duration of the day. This exploratory analysis can be used to estimate the number of pickups in a borough in a particular time period and divert a greater number of Uber cars in the boroughs where the demand is high whereas lower the diversion of the number of Uber cars in the areas with lower demand. The hourly analysis can help the drivers move efficiently in the city as per the peak hour demand in the boroughs. The results may vary in case of very irregular weather conditions.

The dataset contains 9 variables as below:

- pickup_dt: Time period of the observations.
- borough: Regions in NYC.
- pickups: Number of pickups for the period.
- spd: Wind speed in miles/hour.
- vsb: Visibility in Miles to the nearest tenth.
- temp: temperature in Fahrenheit.
- pcp24: 24-hour liquid precipitation.
- sd: Snow depth in inches.
- hday: Being a holiday (Y) or not (N).

3. Method of Sampling

We have taken samples for 2 weekdays, 1 holiday and 1 weekend the months of January, February, May and June, and samples of 2 weekdays and 2 weekends for the months of March and April. The time slots chosen are 8-10 AM which are peak office rush hours, 6-8 PM during which people travel from their workplace back home, and 1-3 AM when people might travel less or not travel at all. These timings are chosen as they tend to have the highest and lowest probability of booking private cab services as per the rush hours. We will do this sampling for each of the 5 boroughs for all 6 months and draw inferences based on the effect of weather on the number of Uber pickups.

The program is made in a way to make it highly scalable and adaptable if any modifications occur. The program is coded in Python and is executed in Jupyter Notebook application. The population was stored as a data frame and further subdivided into smaller data frames based on the time slots and days chosen per month. These smaller data frames are classified as follows:

- Data frame of all months and all boroughs
    - allmallb

- Data frames per month for all boroughs
    - jantestings (data for the month of January)
    - febtest (data for the month of February)
    - martest (data for the month of March)
    - aprtest (data for the month of April)
    - maytest (data for the month of May)
    - juntest (data for the month of June)

- Data frames for all boroughs per month
    - bronxall (data for all months in Bronx)
    - queensall (data for all months in Queens)
    - statall (data for all months in Staten Island)
    - manall (data for all months in Manhattan)
    - broall (data for all months in Brooklyn)

These data frames were then extracted from the program as individual CSV files to run descriptive and inferential statistical analyses on Minitab software. Link to code - https://tinyurl.com/t5jjtvz

4. Descriptive Statistical Analysis

Figure 1 depicts the relationship between temperature (in degrees Fahrenheit) and the number of pickups for each borough of New York from January to June 2015. It can be observed that Manhattan has the highest number of pickups while Staten Island has the least number of pickups. The highest number of pickups go up to 6392 at a temperature of 22 degrees Fahrenheit.

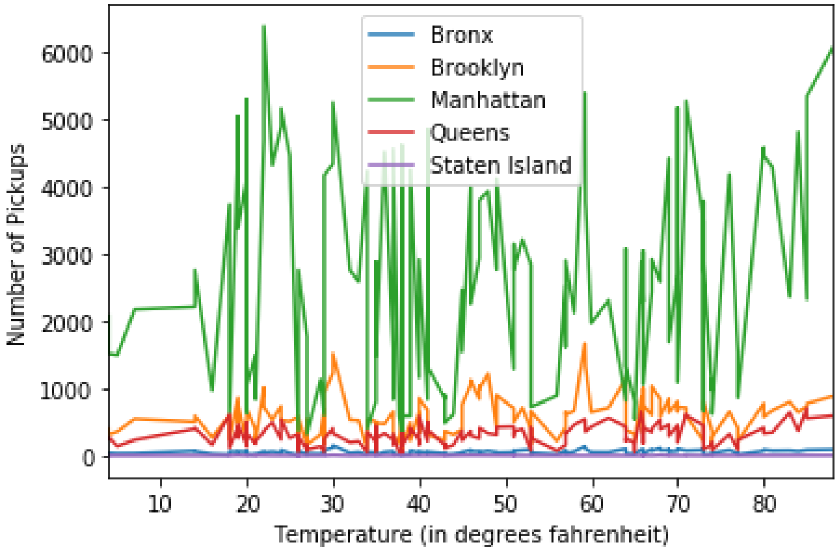
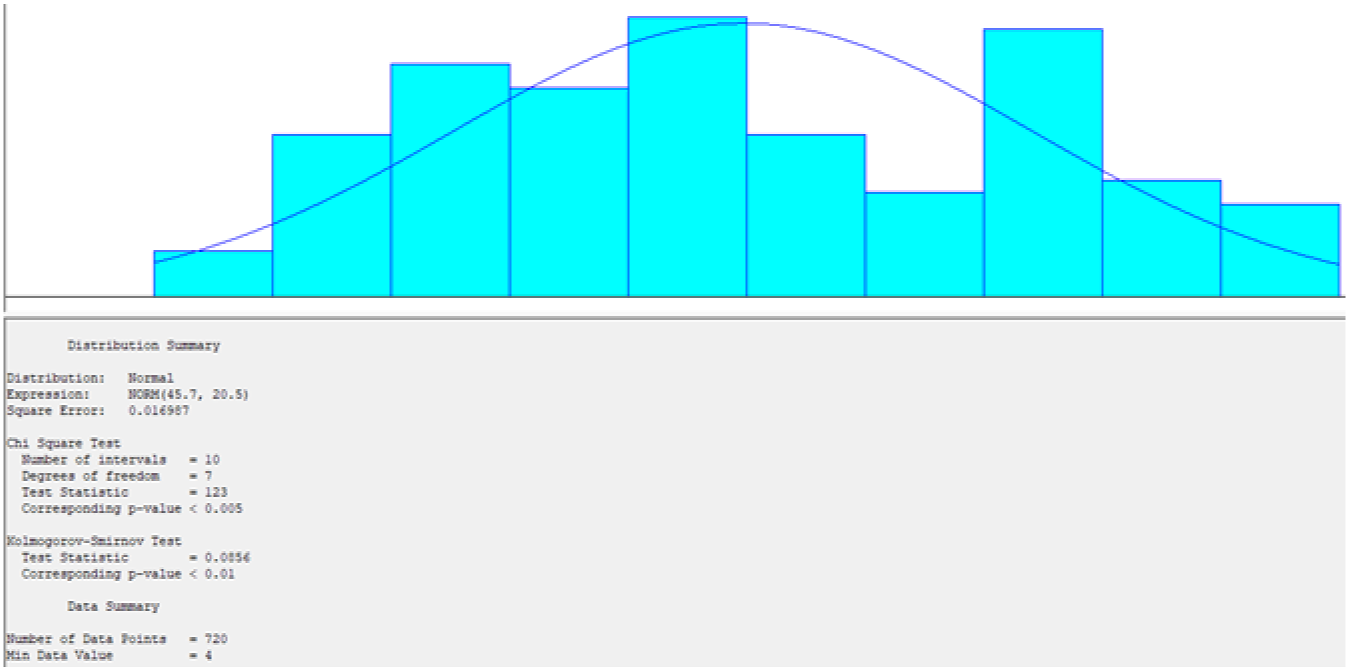


Figure 1: Temperature v/s Number of Pickups of all boroughs



```
        Distribution Summary

Distribution:   Normal
Expression:     NORM(45.7, 20.5)
Square Error:   0.016987

Chi Square Test
  Number of intervals   = 10
  Degrees of freedom    = 7
  Test Statistic        = 123
  Corresponding p-value < 0.005

Kolmogorov-Smirnov Test
  Test Statistic        = 0.0856
  Corresponding p-value < 0.01

        Data Summary

Number of Data Points   = 720
Min Data Value          = 4
```

Figure 2: Normal Distribution Curve for Temperature

Figure 2 shows the normal distribution curve for the Temperature data points in our sample set. We have assumed the distribution to be normal and have carried out the calculations accordingly.

Data for individual boroughs is analyzed for all the given months.

Bronx

Variable	**Mean**	**StDev**	**Variance**	**Minimum**	**Q1**	**Median**	**Q3**	**Maximum**	**Range**	**IQR**
pickups	52.00	28.03	785.76	7.00	28.00	50.00	71.50	152.00	145.00	43.50
temp	45.72	20.57	423.32	4.00	29.00	42.00	65.00	88.00	84.00	36.00

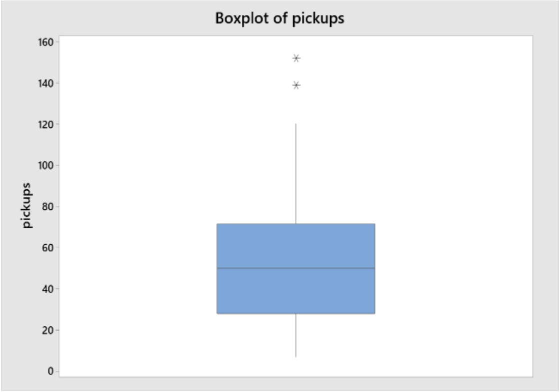


Figure 3.1: Boxplot of Temperature in Bronx

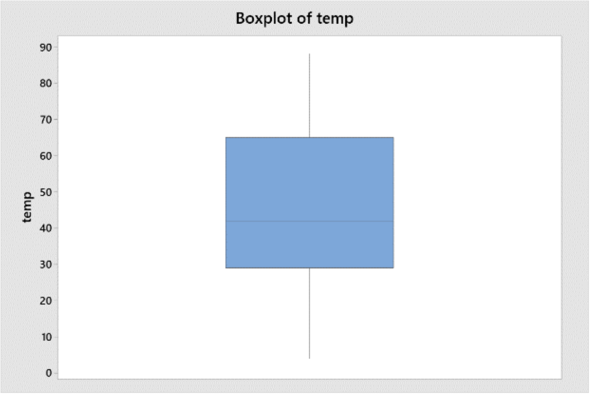


Figure 3.2: Boxplot of pickups in Bronx

Brooklyn

Variable	Mean	StDev	Variance	Minimum	Q1	Median	Q3	Maximum	Range	IQR
pickups	570.9	290.6	84447.3	93.0	342.3	560.5	732.0	1670.0	1577.0	389.8
temp	45.72	20.57	423.32	4.00	29.00	42.00	65.00	88.00	84.00	36.00

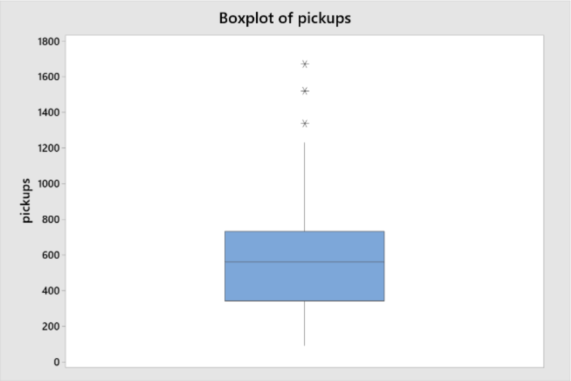


Figure 4.1: Boxplot of Temperature in Brooklyn

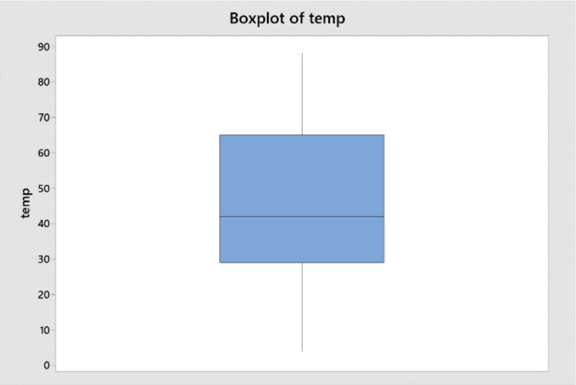


Figure 4.2: Boxplot of Pickups in Brooklyn

Manhattan

Variable	Mean	StDev	Variance	Minimum	Q1	Median	Q3	Maximum	Range	IQR
pickups	2535	1539	2367142	319	1068	2506	3799	6392	6073	2731
temp	45.72	20.57	423.32	4.00	29.00	42.00	65.00	88.00	84.00	36.00

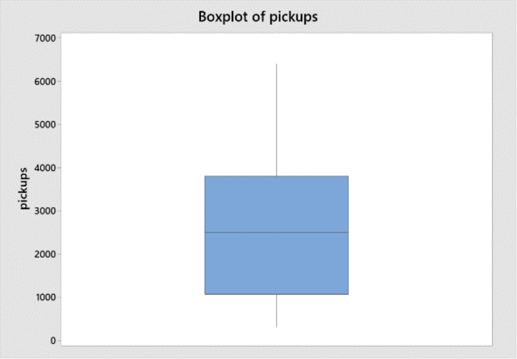


Figure 5.1: Boxplot of Temperature in Manhattan

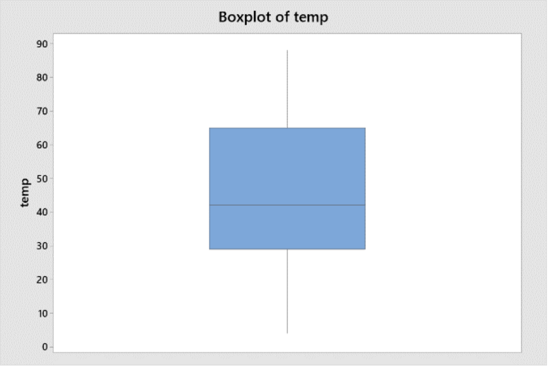


Figure 5.2: Boxplot of Pickups in Manhattan

Queens

Variable	Mean	StDev	Variance	Minimum	Q1	Median	Q3	Maximum	Range	IQR
pickups	302.8	149.6	22385.1	32.0	195.0	300.0	404.0	707.0	675.0	209.0
temp	45.72	20.57	423.32	4.00	29.00	42.00	65.00	88.00	84.00	36.00

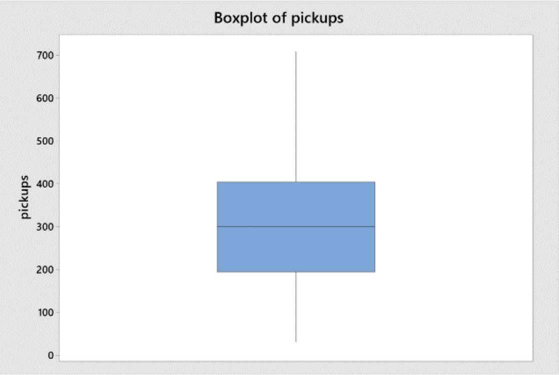


Figure 6.1: Boxplot of Temperature in Queens

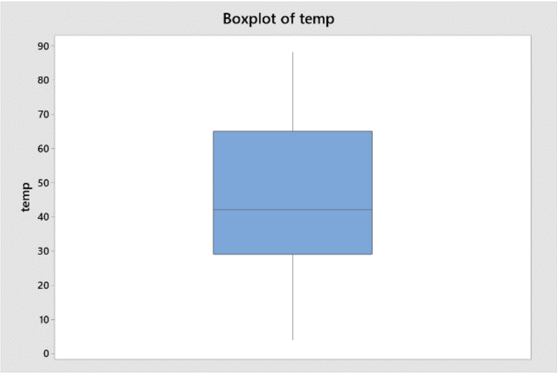


Figure 6.2: Boxplot of Pickups in Queens

Staten Island

Variable	**Mean**	**StDev**	**Variance**	**Minimum**	**Q1**	**Median**	**Q3**	**Maximum**	**Range**	**IQR**
pickups	1.708	1.625	2.642	0.000	0.000	1.000	3.000	7.000	7.000	3.000
temp	45.72	20.57	423.32	4.00	29.00	42.00	65.00	88.00	84.00	36.00

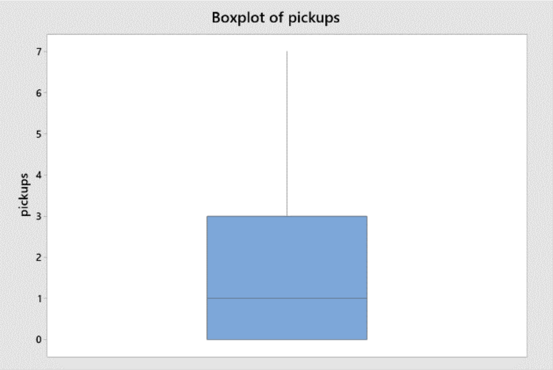


Figure 7.1: Boxplot of Temperature in Staten Island

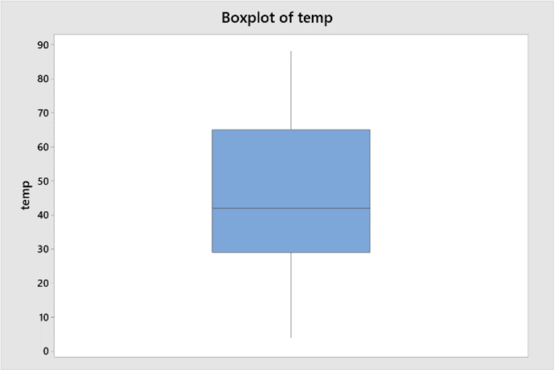


Figure 7.2: Boxplot of Pickups in Staten Island

5. Inferential Statistical Analysis

Test for Equal Variances: Pickups(8-10AM), Pickups(6-8PM) using ANOVA

Method

Null hypothesis	All variances are equal
Alternative hypothesis	At least one variance is different
Significance level	α = 0.05

F method is used. This method is accurate for normal data only.

95% Bonferroni Confidence Intervals for Standard Deviations

Sample	**N**	**StDev**	**CI**
Pickups (8-10AM)	240	852.55	(773.08, 949.47)
Pickups (6-8PM)	240	1562.60	(1416.94, 1740.23)

Individual confidence level = 97.5%

Tests

Method	Test Statistic	P-Value
F	0.30	0.000

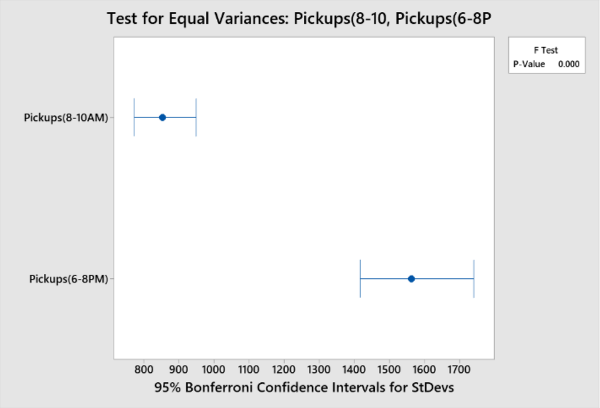


Figure 8: Test for Equal Variances (8-10 am v/s 6-8 pm pickups)

Two-Sample T-Test and Confidence Interval: Pickups(8-10AM), Pickups(6-8PM)

Method

μ_1: mean of Pickups(8-10AM)
μ_2: mean of Pickups(6-8PM)
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Pickups(8-10AM)	240	543	853	55
Pickups(6-8PM)	240	1018	1563	101

Estimation for Difference

Difference	95% Lower Bound for Difference
-475	-664

Hypothesis Testing

Null hypothesis	H_0: $\mu_1 - \mu_2 = 0$
Alternative hypothesis	H_1: $\mu_1 - \mu_2 > 0$

T-Value	DF	P-Value
-4.13	369	1.000

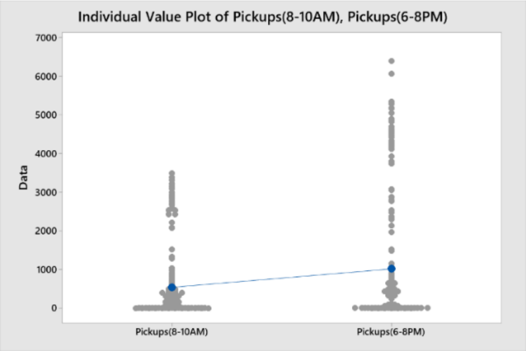


Figure 4: Individual Value plots of Pickups (8-10 am v/s 6-8 pm)

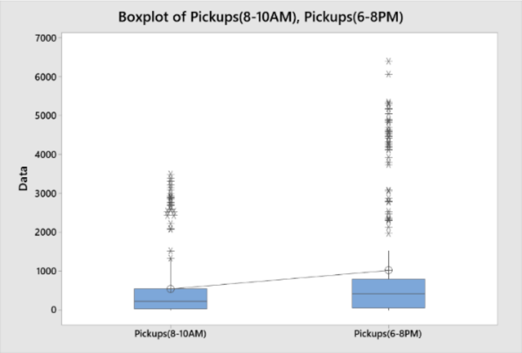


Figure 3: Boxplot of Pickups (8-10 am v/s 6-8 pm)

Test and CI for One Proportion of Snow Depth (sd)

No. of success cases in the proportion are all data points which are greater than 0.054 i.e. snow depth greater than 0.054

Method

p: event proportion
Normal approximation method is used for this analysis.

Descriptive Statistics

N	Event	Sample p	95% CI for p
720	130	0.180556	(0.152459, 0.208652)

Correlation Between Pickups and Snow Depth

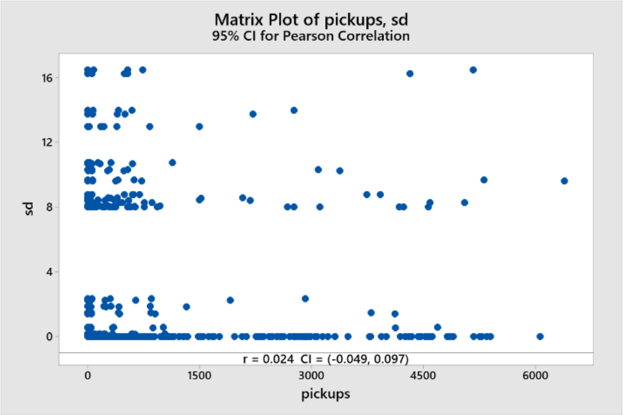


Figure 5: Correlations Coefficient using Matrix Plot for Pickups and Snow Depth

Method

Correlation type: Pearson
Number of rows used: 720

ρ: pairwise Pearson correlation

Correlations

	Pickups
sd	0.024

Pairwise Pearson Correlations

Sample 1	Sample 2	N	Correlation	95% CI for ρ	P-Value
sd	pickups	720	0.024	(-0.049, 0.097)	0.518

6. Conclusion

We have assumed a normal Temperature distribution for all the 5 boroughs of New York for all the 6 months for which the data set is used. Descriptive Statistical Analysis was carried out for the number of pickups in each borough against the temperature from January to June. It was commonly observed in all boroughs that the number of pickups was more at lesser temperatures and during evening peak hours (6-8 pm). Box plots for individual pickups and temperatures in each borough for the six months from January to June were plotted and displayed in this report which shed light on the individual data set distributions.

The correlation coefficient for the Snow Depth and the number of pickups gave us information about the trends followed in the number of pickups due to the snowfall and the depth of snow. It is logical to conclude that the snow depth does not majorly affect the number of Uber pickups in New York.

Furthermore, test for equal variances was carried out using ANOVA in Minitab, which concluded that the variances for the 2 samples (pickups during 8-10 am and pickups during 6-8 pm) were significantly different. We also found out the T-values and Confidence Intervals for both the samples, which was then used for the Hypothesis Testing for the difference of means. By conducting the Hypothesis testing, we concluded that the difference of means for the above mentioned 2 samples is greater than 0 and it is not the same for those samples. We rejected the null hypothesis which stated that the difference between the means of the 2 samples is zero. The mean of the number of pickups during 6-8 pm was significantly higher than that of the pickups during 8-10 am for all the boroughs for all the months. Hence, Uber would have to keep a greater number of cabs available during 6-8 pm in all the given months because of the high demand during that time considering all weather conditions.

7. Technologies Used

- MS Excel software
- Minitab software
- Python programming language
- Jupyter Notebook application (used for running the python program)
- Arena software

8. References

- Pappas, Y. (2017, February 13). NYC Uber Pickups with Weather and Holidays. Retrieved from https://www.kaggle.com/yannisp/uber-pickups-enriched