Problem Set - Uber Case

AUTHOR
Prateek Naharia

Install the packages if you don't have them installed yet. Install them only once, and not when you run you render.

1. Read the HBS Case. What is the difference between Uber POOL and Express POOL? No more than two sentences.

The major difference is Express POOL offered cheaper rides with walking and waiting up-to 2-5 minutes for more efficient matches to have higher seat occupancy.

2. How did Uber use surveys in designing Uber Express Pool?

Uber used conjoint surveys(type of questionnaire) to gather data that measured consumers sensitivity in terms of their service to understand their preferences and willingness to walk and wait for a ride & sensitivity to pricing (floor & ceiling).

3. Suppose Uber was considering a new algorithm to recommend ride destinations in the app. Which type of research strategy should they use (A/B Test, Switchback, Synthetic Control)? No more than two sentences.

Should consider A/B test to evaluate the effectiveness of a new algorithm for recommending ride destinations in the app, with one group receiving the new algorithm and the other group serving as a control group.

4. Suppose Uber was considering a radio advertising campaign. Which type of research strategy should they use (A/B Test, Switchback, Synthetic Control)? No more than two sentences.

Synthetic control is an effective research method that allows for the measurement of the treatment effect between similar groups and can provide reliable results. Therefore, it could be used to evaluate the impact of an intervention or policy change.

5. Create two new columns in the dataset that represent the total number of trips for both pool products and the profit from these products. (10 points)

(remember you can create a new column by: data[, new_col_name := whatever you want the new column to contain])

```
data[, totaltrips := trips_express_pool+trips_pool]
data[, profit := revenue-total_driver_payout_sr]
head(data)
```

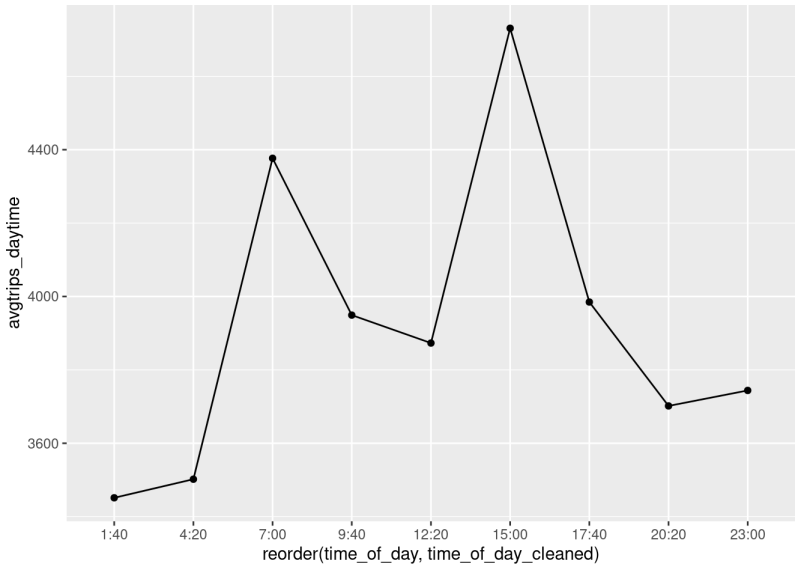
```
   city_id  period_start wait_time treat commute trips_pool trips_express_pool
1:  Boston  2/19/18 7:00    2 mins FALSE    TRUE       1417               3252
2:  Boston  2/19/18 9:40    5 mins  TRUE   FALSE       1462               2364
3:  Boston 2/19/18 12:20    2 mins FALSE   FALSE       1360               2189
4:  Boston 2/19/18 15:00    5 mins  TRUE    TRUE       1977               3580
5:  Boston 2/19/18 17:40    2 mins FALSE   FALSE       1368               2575
6:  Boston 2/19/18 20:20    5 mins  TRUE   FALSE       1404               2027
   rider_cancellations total_driver_payout_sr total_matches_sr
1:                 256                 34459             3365
2:                 201                 29770             2292
3:                 115                 27446             2288
4:                 356                 44992             4040
5:                 187                 27583             2193
6:                 133                 23892             2070
   total_double_matches_sr revenue totaltrips profit
1:                    1479   46042       4669  11582
2:                    1279   32174       3826   2404
3:                     957   31043       3549   3596
4:                    2029   53747       5557   8755
5:                     975   32800       3943   5217
6:                    1069   29291       3431   5399
```

6. Plot the average number of trips as a function of the time of the day. Describe a reason why this pattern exists (no more than 2 sentences). (20 points)

Hint: You can use ggplot to do this. As in assignment 1, you'll first have to create a dataset with the average number of trips by time of the day.

```
library(ggplot2)
data[, time_of_day := str_split_fixed(period_start, " ", n=2)[, 2]] #spliting date-time
data[, date_start := str_split_fixed(period_start, " ", n=2)[, 1]]
plot_data <- data[, list(avgtrips_daytime = mean(totaltrips)), by=.(time_of_day)]
plot_data[, time_of_day_cleaned := as.integer(str_replace_all(time_of_day, ":", ""))]
plot_data <- plot_data[order(time_of_day_cleaned),]


# Plotting the line graph
this_plot <- ggplot(plot_data, aes(
x = reorder(time_of_day, time_of_day_cleaned),y = avgtrips_daytime)
) + geom_point() + geom_line(aes(group=1))
this_plot
```



The above plot interpretation is as, It shows peaks in the number of average trips during the expected rush hour time frames of 7-9:40 AM and 15:00-1740 hours, with more number of trips, which matches the information mentioned in the reading about when rush hours typically occur.

7. Conduct a regression analysis of the experiment (considering the outcomes: revenue, total_driver_payout_sr, rider_cancellations, total_trips). Make sure to think carefully about the correct regression specification. The regression output should be easy to read, so use 'etable' or 'modelsummary'. What do you learn in words from this regression analysis (no more than 5 sentences but it can be less)?

Hint: We should control for the fact that different times of the day and different days have different demand patterns. (Please refer to p.13 of the HBS article to see why) Hint: The syntax for fixed effects is: feols(outcome ~ treatment_name | fixed_effect_name1 + fixed_effect_name2, data = data, se = 'hetero') Hint: You can output multiple regressions in this way: etable(reg1, reg2)

```
data[, wait_time := as.integer(str_sub(wait_time, start = 1, end = 2))]
data[, date_start := as.integer(str_replace_all(date_start, "/", ""))]
data[, commute := as.integer(commute)]
data[, treat := as.integer(treat)]
# DATE, HOUR, WEEK EXTRACTION
data[, period_start := as.POSIXct(period_start, format = "%m/%d/%y %H:%M")]
data[, hour_of_day := as.numeric(format(period_start, "%H"))]
data[, day_of_week := as.factor(weekdays(period_start))]
```

Regression considering the outcomes: revenue, total_driver_payout_sr, rider_cancellations, total_trips

```
regression <-  feols(
    c(revenue,total_driver_payout_sr,rider_cancellations, totaltrips) ~ treat
    | day_of_week + hour_of_day, data = data, se = 'hetero'
  )

etable(regression)
```

```
                 regression.1           regression.2        regression.3
Dependent Var.:       revenue total_driver_payout_sr rider_cancellations

treat           -272.1 (878.2)     -2,106.8** (744.7)    26.52*** (6.638)
Fixed-Effects:  --------------- ---------------------- --------------------
day_of_week                Yes                    Yes                 Yes
hour_of_day                Yes                    Yes                 Yes
_______________ _______________ ______________________ ____________________
S.E. type       Heterosk.-rob. Heteroskedastici.-rob. Heteroskedast.-rob.
Observations               126                    126                 126
R2                     0.56975                0.48275             0.55182
```

```
Within R2              0.00087              0.06783              0.12674

                   regression.4
Dependent Var.:     totaltrips

treat            -88.37 (77.30)
Fixed-Effects:   --------------
day_of_week                 Yes
hour_of_day                 Yes
_________________ ______________
S.E. type         Heterosk.-rob.
Observations                 126
R2                       0.49501
Within R2                0.01174
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is observed from the above regression that - Treatment (treat) is -272.1(-272.1 (878.2)) & -2106.8 (-2,106.8** (744.7)) is negative for revenue & total_driver_payout_sr. While 26.52*** (6.638) for rider_cancellations is positive. We can also conclude that total_driver_payout_sr & rider_cancellations are statistically significant. while the other two revenue & total trips are not.

According to the analysis, implementing a 5 minute wait period has a positive impact on Uber's driver payout compared to a 2 minute wait period. However, this policy has negative effects on other aspects of Uber's operations, such as reducing revenue, increasing rider cancellations, and decreasing the overall number of trips taken.

8. One of your data scientists suggests that the optimal wait time may differ by whether it's a commuting period. Test whether the effects of a 5 minute wait period on total trips and cancelations differ by whether it's a commuting period (the column 'commute'). Which policy works better during commute times? (10 points)

```
#Performing regression for totaltrips with treat  & commute, & rider cancellations & treat

#totaltrips ~ treat * commute
regression_totaltrips = feols(totaltrips ~ treat * commute,
data = data, se = 'hetero')
#rider_cancellations ~ treat * commute
regression_ridercancellations = feols(rider_cancellations ~ treat * commute,
data = data, se = 'hetero')

#summary for both the regressions total_trips & ridercancellations is as below:-

etable(regression_totaltrips, regression_ridercancellations)
```

```
                regression_total.. regression_riderc..
Dependent Var.:         totaltrips rider_cancellations

Constant        3,764.7*** (49.04)    149.1*** (3.673)
treat               -44.28 (71.46)    20.81*** (4.813)
commute         1,280.6*** (160.3)    96.31*** (16.27)
treat x commute     -277.7 (229.5)      35.99. (18.32)
_______________ __________________ ___________________
S.E. type       Heteroskedas.-rob. Heteroskedast.-rob.
Observations                   126                 126
R2                         0.54891             0.72625
Adj. R2                    0.53782             0.71952
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is observed that the interaction effect between the treatment and commute variables is not significant in both regression models. It is also seen that treatment with waiting for 5 minutes has a negative impact during rush hours, hence increasing no of passenger/rider cancellations & decreasing total trips. For commute time - 2 minutes policy works best as wait period.