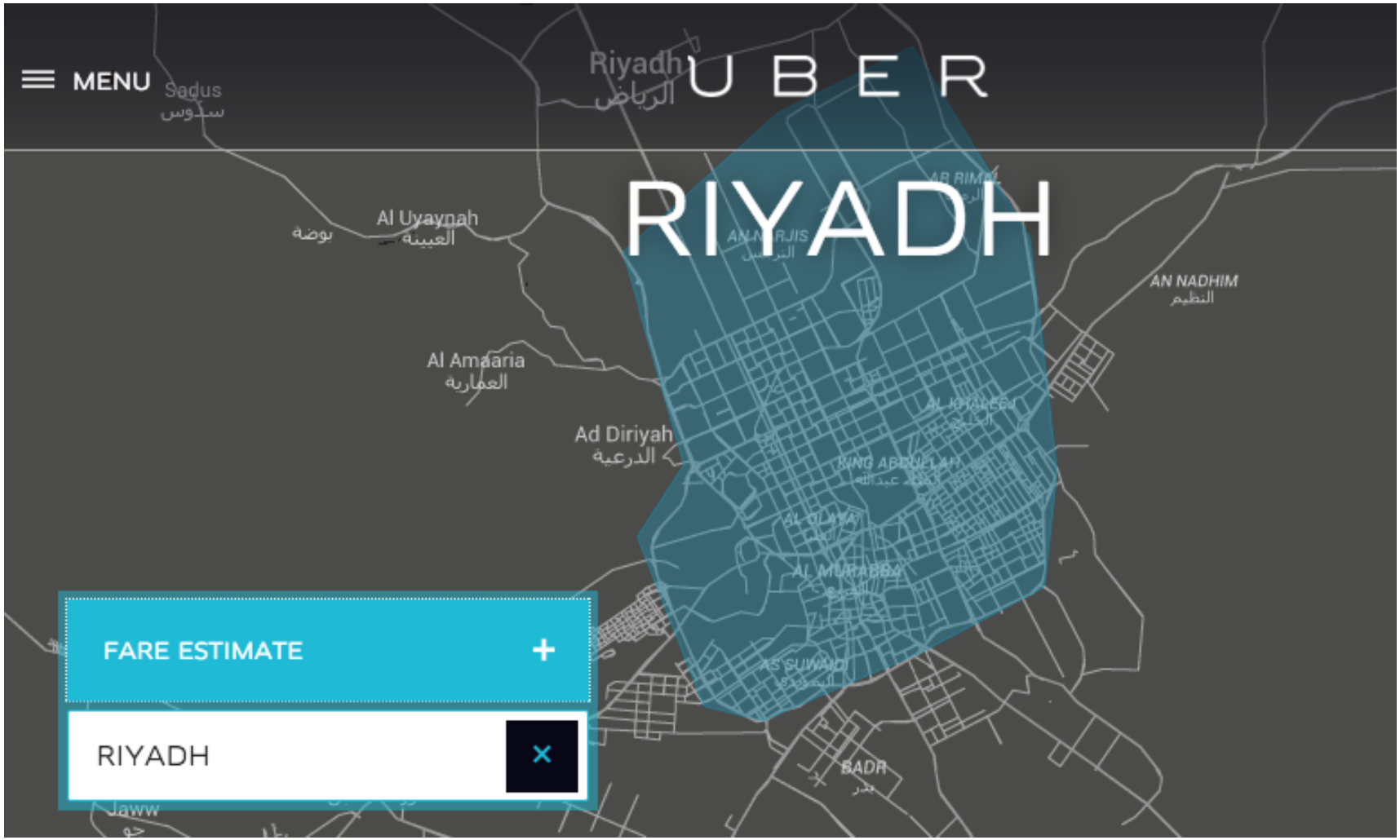
Uber - Internal Audit Analytics Exercise: 2020 strategy for Riyadh Market

Out[1]: The raw code for this IPython notebook is by default hidden for easier reading. To toggle on/off the raw code, click here.

Out[2]:



Introduction

Uber Technologies Inc. is investing $250 million to expand in the Middle East and North Africa, which have some of the ride-sharing service's fastest-growing markets, Bloomberg reports.

Uber is already in Saudi Arabia, and the ride-sharing app is having a significant impact on the transportation economy there.

Problem Statement:

- Understanding the needs of key stakeholders and performing analysis/prototyping solutions
- In this exercise, EMEA rideshare data is to be analyzed and use it to draw a conclusion for the 2020 strategy for the Riyadh market.

Reading, Cleaning and Compiling Data from the given CSV data sample

Importing Libraries

The following libraries should be imported to run this notebook: pandas, sqlite3, numpy, matplotlib, plotly, dash, pivottablejs.

The Plotly Python library is an interactive open-source library. This can be a very helpful tool for data visualization and understanding the data simply and easily. plotly graph objects are a high-level interface to plotly which are easy to use. It can plot various types of graphs and charts like scatter plots, line charts, bar charts, box plots, histograms, pie charts, etc.

Dash is a Python framework for building analytical web applications. Dash helps in building responsive web dashboards that is good to look at and is very fast without the need to understand complex front-end frameworks or languages such as HTML, CSS, JavaScript.

PivotTable.js is a Javascript Pivot Table and Pivot Chart library with drag'n'drop interactivity, and it can now be used with Jupyter/IPython Notebook via the pivottablejs module.

Out[4]: The raw code for this IPython notebook is by default hidden for easier reading. To toggle on/off the raw code, click here.

Exploratory Data Analysis

Exploratory Data Analysis or (EDA) is understanding the data sets by summarizing their main characteristics often plotting them visually.Through the process of EDA, we can ask to define the problem statement or definition on our data set which is very important.

Units considered to perform the analysis/output:

1. Distance: miles
2. Time: Local time in minutes
3. Fare: USD

Displays datatype of all columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40000 entries, 0 to 39999
Data columns (total 36 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   pickup_local_time      28750 non-null  object 
 1   pickup_utc_time        28750 non-null  object 
 2   cancel_fee_local       40000 non-null  float64
 3   cancel_fee_usd         40000 non-null  float64
 4   city_id                40000 non-null  int64  
 5   rider_app              39846 non-null  object 
 6   rider_device           40000 non-null  object 
 7   rider_trip_count       28556 non-null  float64
 8   rider_id               40000 non-null  object 
 9   partner_vehicle_count  40000 non-null  int64  
 10  driver_trip_count      28556 non-null  float64
 11  driver_id              40000 non-null  object 
 12  dropoff_local_time     28556 non-null  object 
 13  dropoff_utc_time       28556 non-null  object 
 14  esttime_to_pickup      38529 non-null  float64
 15  request_type           38537 non-null  object 
 16  entered_destination    40000 non-null  bool   
 17  paid_cash              40000 non-null  bool   
 18  completed_trip         40000 non-null  bool   
 19  surged_trip            40000 non-null  bool   
 20  trip_fare_local        40000 non-null  float64
 21  trip_fare_usd          40000 non-null  float64
 22  partner_id             40000 non-null  object 
 23  request_local_time     40000 non-null  object 
 24  request_utc_time       40000 non-null  object 
 25  distance_to_pickup     38601 non-null  float64
 26  time_to_pickup         28750 non-null  float64
 27  trip_status            40000 non-null  object 
 28  trip_distance_miles    39999 non-null  float64
 29  trip_duration_seconds  40000 non-null  int64  
 30  trip_id                40000 non-null  object 
 31  vehicle_trip_count     28556 non-null  float64
 32  vehicle_id             40000 non-null  object 
 33  vehicle_type           39693 non-null  object 
 34  pickup_geo             40000 non-null  object 
 35  dropoff_geo            40000 non-null  object 
```

```
dtypes: bool(4), float64(11), int64(3), object(18)
memory usage: 9.9+ MB
```

Displays first five rows of the dataset

Out[9]:

	pickup_local_time	pickup_utc_time	cancel_fee_local	cancel_fee_usd	city_id	rider_app	rider_device	rider_trip_count	rider_id
0	2018-05-10 09:00:00	2018-05-10 06:00:00	0.0	0.0	1	3.298.10000	iphone	131.0	888bb3c7-c55b-5d41-8cd9-9a4a2554b4e4
1	NaN	NaN	0.0	0.0	1	3.298.10000	iphone	NaN	cfc5db3c-e25e-5f48-9d59-8f57fb611f86
2	2018-05-17 09:00:00	2018-05-17 06:00:00	0.0	0.0	1	3.268.10002	iphone	53.0	5296a57d-3294-54a7-a0a0-e7fa7e6d8caa
3	NaN	NaN	0.0	0.0	1	3.275.10002	iphone	NaN	0792af9c-547c-56ed-adce-e217b7eed8d2
4	2018-05-10 16:00:00	2018-05-10 13:00:00	0.0	0.0	1	3.241.2	iphone	499.0	7aebd941-f808-54e5-a7ce-ae890ed6e1e8

5 rows × 36 columns

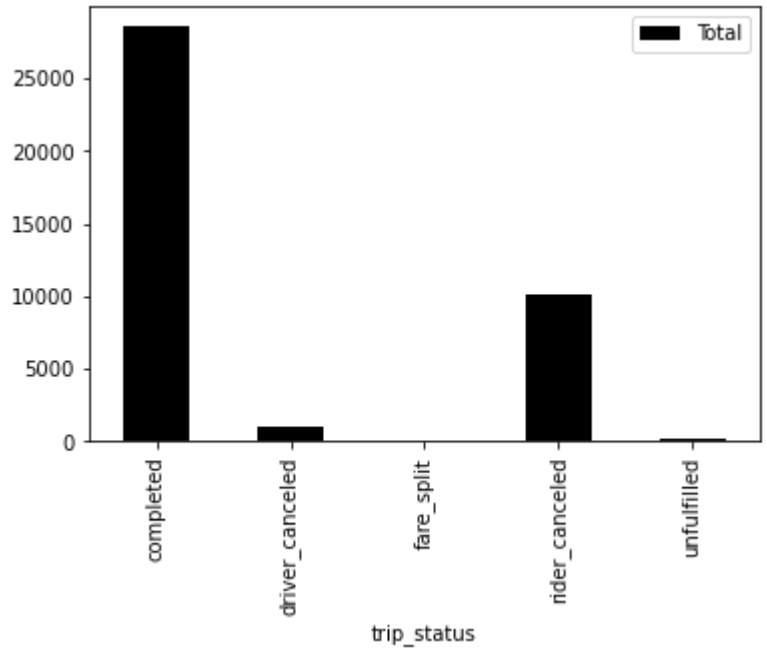
Part 2: Analysis and Presentation

Statistical information about dataset

Out[16]:

	cancel_fee_local	cancel_fee_usd	city_id	rider_trip_count	partner_vehicle_count	driver_trip_count	esttime_to_pickup	trip_fare_lo
count	40000.000000	40000.000000	40000.0	28556.000000	40000.000000	28556.000000	38529.000000	40000.0000
mean	0.209350	0.055824	1.0	170.142912	3.248325	1237.466452	282.948896	16.6363
std	1.293074	0.344802	0.0	250.283997	9.602735	1455.880407	180.717438	18.0275
min	0.000000	0.000000	1.0	1.000000	1.000000	1.000000	1.000000	0.0000
25%	0.000000	0.000000	1.0	22.000000	1.000000	251.000000	166.000000	0.0000
50%	0.000000	0.000000	1.0	77.000000	1.000000	693.000000	253.000000	12.0000
75%	0.000000	0.000000	1.0	212.000000	1.000000	1652.250000	361.000000	23.5200
max	20.000000	5.333483	1.0	3166.000000	73.000000	12735.000000	4073.000000	440.7000

Out[17]: <AxesSubplot:xlabel='trip_status'>



Below are some numerical findings from the data which will further help to analyze data

Completion Rate of rides

```
Completion_rate: 71.37%
```

Cancellation Rate of rides

```
Cancellation_rate: 28.63%
```

Total fare of rides

```
Total_Fare: $ 177445.52
```

Percentage of Surged trips

```
Surged_trip_pct:25.42%
```

Average fare of completed rides

```
Average_Fare: $ 6.22
```

Trip Fare on basis of Surge

Out[31]:

	surged_trip	trip_fare_usd
0	No Surge	5.531958
1	Surge	8.036503

Overview of the data:

1. The dataset provided gives information of 40k trips in the EMEA region – Riyadh city
2. Number of trips:
   - Completed: 28.5k (71.37%)
   - Trips cancelled (rider): 10.1k (25.43%)
   - Trips cancelled (driver): 1k (2.5%)
   - Trips unfulfilled: 234 (0.58%)
   - Fare Split: 8 (0.02%)
3. Average fare/ride: $6(88\%paidincash), total fare$:177k (completed rides)
4. Surged trips: 25% of the completed trips

Inference of the dataset

The below dashboard gives information on the Cancellation rate, Total fare, Average fare, Surged trip percentage. It is a dynamic tool, whenever the data changes output will also change accordingly.

To see the dynamic table, please click on the below link

```
Dash is running on http://127.0.0.1:8050/ (http://127.0.0.1:8050/)

 * Serving Flask app "__main__" (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production deployment.
   Use a production WSGI server instead.
 * Debug mode: off

 * Running on http://127.0.0.1:8050/ (http://127.0.0.1:8050/) (Press CTRL+C to quit)
127.0.0.1 - - [19/Oct/2021 08:28:08] "GET / HTTP/1.1" 200 -
127.0.0.1 - - [19/Oct/2021 08:28:09] "GET /_dash-layout HTTP/1.1" 200 -
127.0.0.1 - - [19/Oct/2021 08:28:09] "GET /_dash-dependencies HTTP/1.1" 200 -
127.0.0.1 - - [19/Oct/2021 08:28:09] "GET /_favicon.ico?v=2.0.0 HTTP/1.1" 200 -
127.0.0.1 - - [19/Oct/2021 08:28:09] "GET /_dash-component-suites/dash/dash_table/async-highlight.js HTTP/1.1"
200 -
127.0.0.1 - - [19/Oct/2021 08:28:09] "GET /_dash-component-suites/dash/dash_table/async-table.js HTTP/1.1" 200
-
```

a. Analysis based on Rider experience

a. Analysis based on Rider experience

The first analysis made is to check the Rider's experience based on given ETA and actual pick up time. From the table below it is observed that the time when the rider gets picked up is, in all cases, greater than the estimated pick up time.

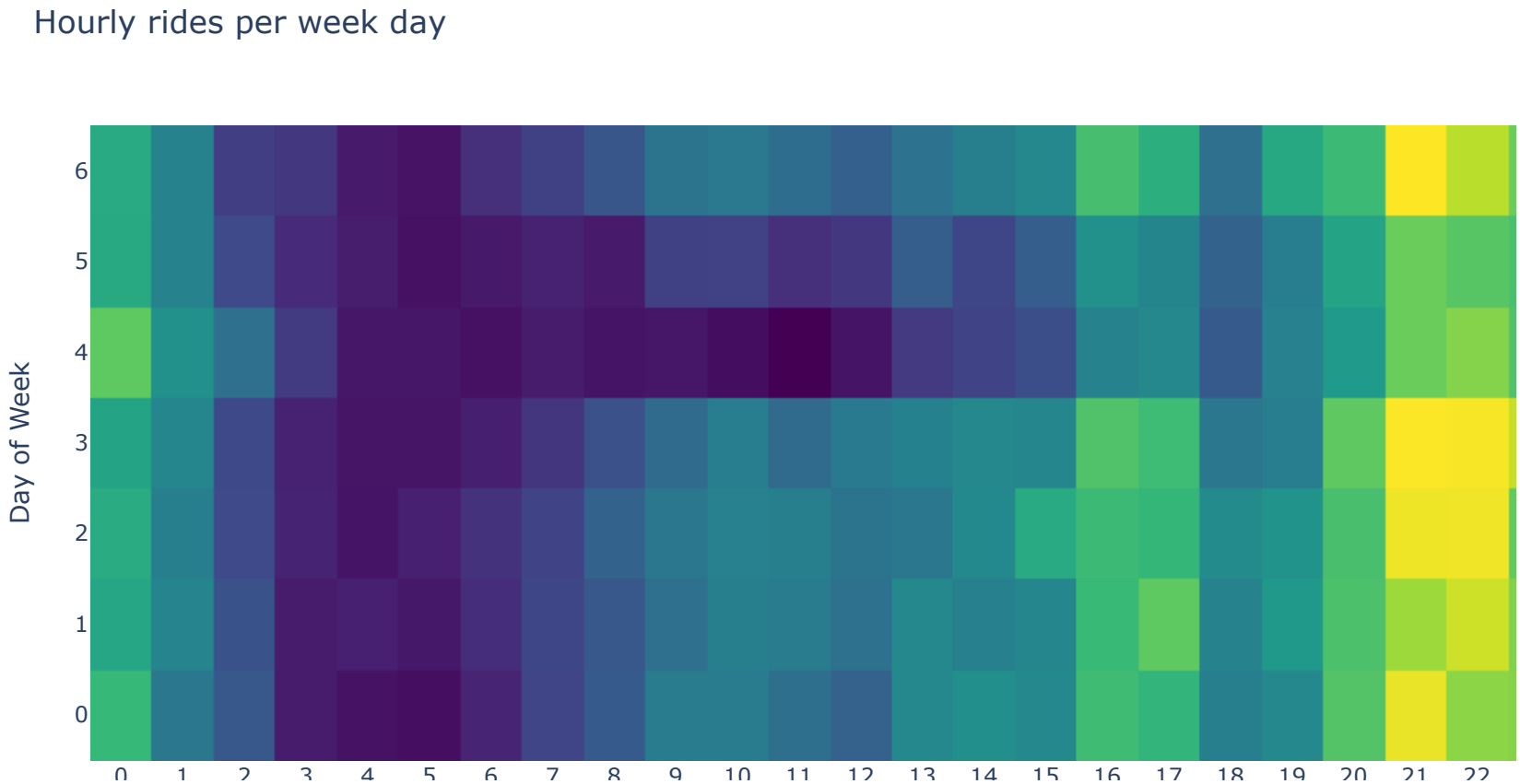
Recommendation: Ensure that the ETA is monitored accurately so that the rider's expectation is set and the driver can try to meet it as well.

Out[38]:

	Distance_pick_up_window	time_to_pickup	esttime_to_pickup
0	0-1	7.967660	4.291548
1	1-2	10.879167	7.091667
2	2-3	13.107143	8.352381
3	3-4	18.472222	9.222222
4	4-5	41.666667	9.183333
5	>5	27.183333	7.003333

b. Analysis based on picked up rides by hours and days of the week

Here, the analysis is done based on the picked up rides by hour and days of the week. It is to be observed from the heatmap that there is a huge spike in rides between 8pm to 11pm almost everyday in the week.



Interactive tool - Picked up rides by hours and days of the week

To see the interactive visualization, please click on the below last link:

```
Dash is running on http://127.0.0.1:8050/ (http://127.0.0.1:8050/)

Dash is running on http://127.0.0.1:8050/ (http://127.0.0.1:8050/)

Dash is running on http://127.0.0.1:8050/ (http://127.0.0.1:8050/)

 * Serving Flask app "__main__" (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production deployment.
   Use a production WSGI server instead.
 * Debug mode: off


 * Running on http://127.0.0.1:8050/ (http://127.0.0.1:8050/) (Press CTRL+C to quit)
127.0.0.1 - - [19/Oct/2021 08:30:16] "GET / HTTP/1.1" 200 -
127.0.0.1 - - [19/Oct/2021 08:30:17] "GET /_dash-layout HTTP/1.1" 200 -
127.0.0.1 - - [19/Oct/2021 08:30:17] "GET /_dash-dependencies HTTP/1.1" 200 -
127.0.0.1 - - [19/Oct/2021 08:30:17] "GET /_dash-component-suites/dash/dcc/async-dropdown.js HTTP/1.1" 200 -
127.0.0.1 - - [19/Oct/2021 08:30:17] "GET /_dash-component-suites/dash/dcc/async-graph.js HTTP/1.1" 200 -
127.0.0.1 - - [19/Oct/2021 08:30:17] "POST /_dash-update-component HTTP/1.1" 200 -
127.0.0.1 - - [19/Oct/2021 08:30:17] "GET /_dash-component-suites/dash/dcc/async-plotlyjs.js HTTP/1.1" 200 -
```

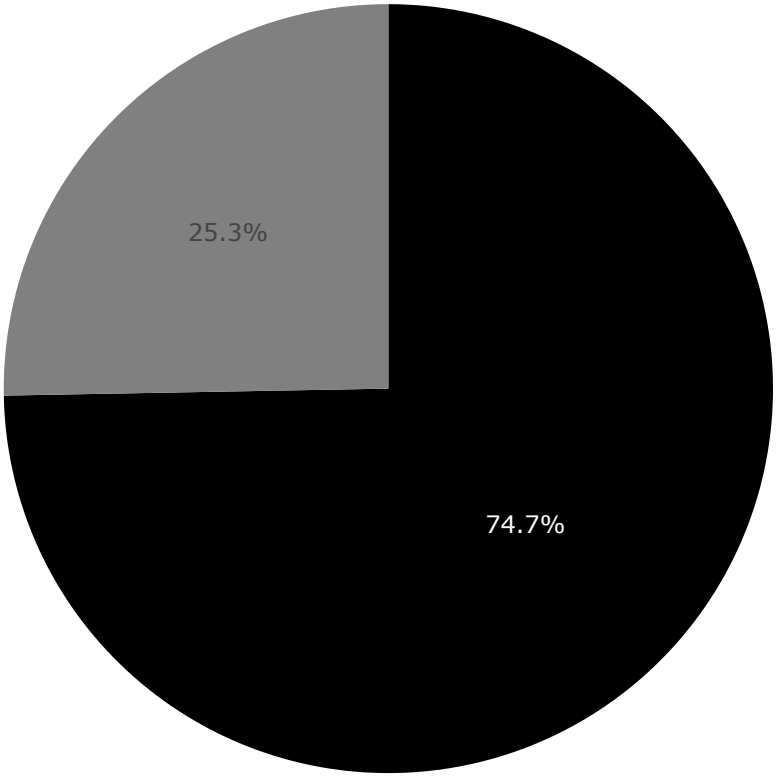
c. Cancellation Analysis on driver canceled rides during Surge

Out[58]:

	surged_trip	trip_id
0	No Surge	774
1	Surge	262

Out[60]:

	surged_trip	Driver_Cancelled_trip	Percent_total_driver_cancelled_trips
0	No Surge	774	74.710425
1	Surge	262	25.289575



Driver cancellation during surge:

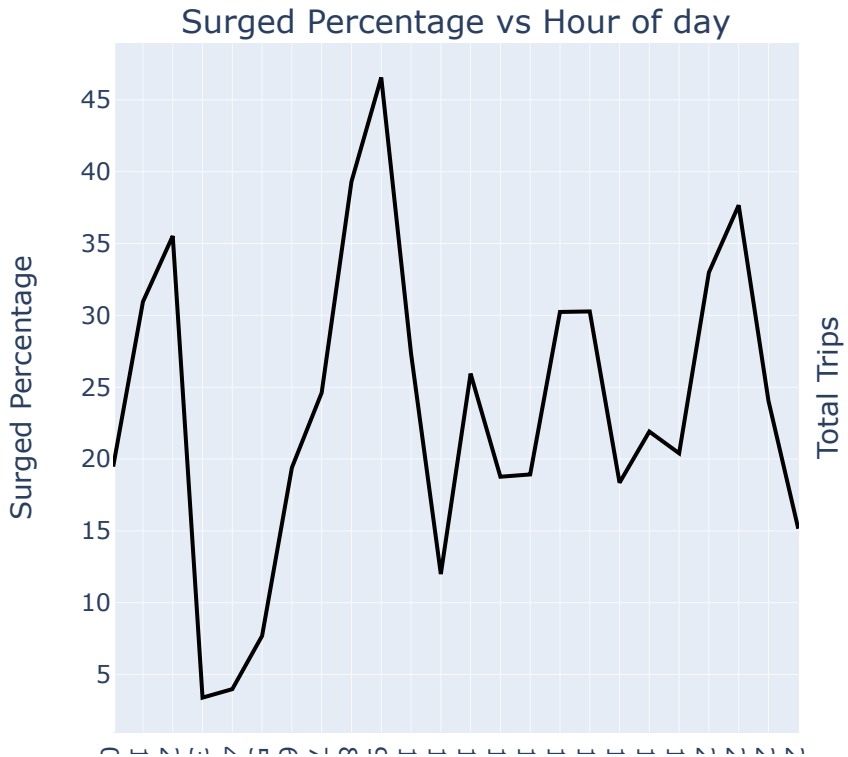
- Demand for rides increases -> Prices go up -> Riders pay more/wait

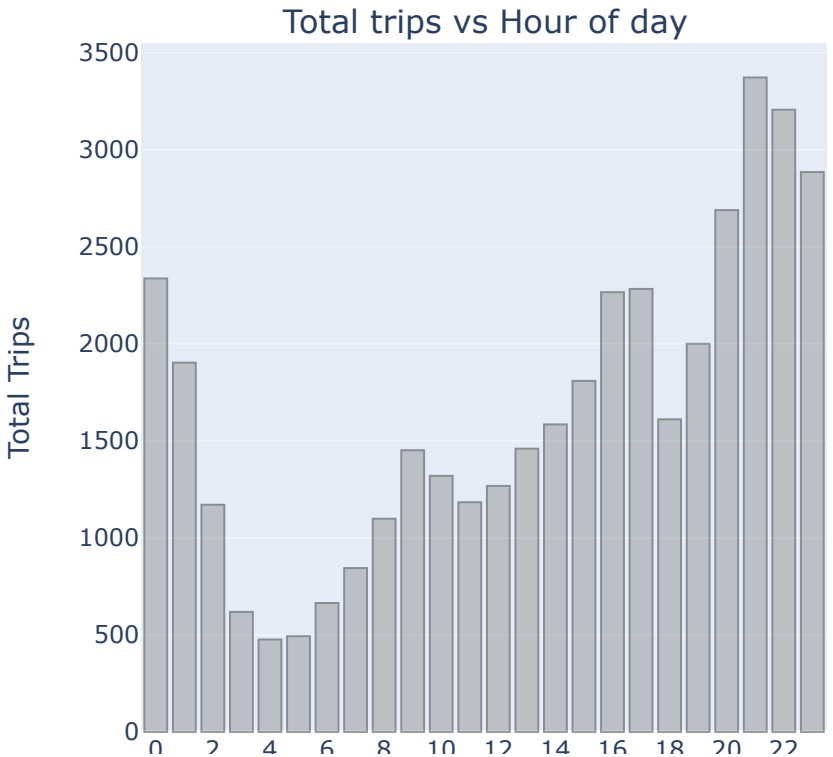
a. **Findings**: Based on the graph, 25% of the trips are getting canceled by the driver during surge

b. **Hypothesis**: Due to high surge pricing, the revenue lost for uber for cancellation of a ride during surge hours will be more than the revenue lost for cancellation of a ride during non-surge hours. If the cancellations can be further reduced to less than 25% then more revenue can be generated.

c. **Recommendation**: Advantages of surge pricing could be explained to drivers and need to ensure that the drivers don't cancel the rides during peak hours. Drivers could be incentivized if the trip is not getting canceled by a driver during surge hours.

d. Surge Analysis based on hour of the day





How does surge works?

- Demand for rides increases -> Prices go up -> Riders pay more/wait

a. **Hypothesis**: Based on the graph, let's assume that the average surge should be at 20%

b. **Findings**: There is a huge spike in rides at 9 AM (28%) and another at 9 PM (17%) – local time considered

c. **Conclusion**: Uber should ensure that during these times, request should be fulfilled to provide a better rider experience

d. **Recommendation**: Advantages of surge pricing could be explained to drivers and need to ensure that the drivers don't cancel the rides during peak hours. Areas that are busiest during peak hours could be allocated with more drivers to ensure that requests are fulfilled

Interactive tool - Surge Analysis

To see the interactive visualization, please click on the below last link:

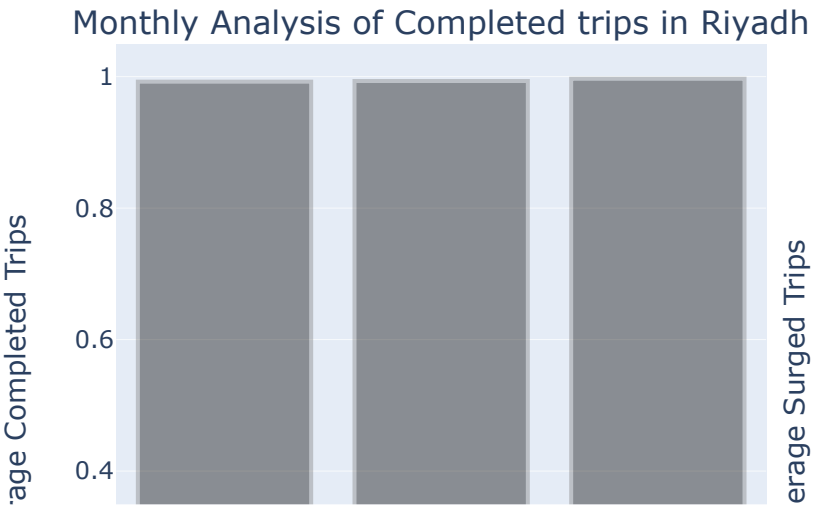
```
Dash is running on http://127.0.0.1:8050/ (http://127.0.0.1:8050/)

Dash is running on http://127.0.0.1:8050/ (http://127.0.0.1:8050/)

Dash is running on http://127.0.0.1:8050/ (http://127.0.0.1:8050/)

Dash is running on http://127.0.0.1:8050/ (http://127.0.0.1:8050/)

Dash is running on http://127.0.0.1:8050/ (http://127.0.0.1:8050/)

 * Serving Flask app "__main__" (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production deployment.
   Use a production WSGI server instead.
 * Debug mode: off

 * Running on http://127.0.0.1:8050/ (http://127.0.0.1:8050/) (Press CTRL+C to quit)
127.0.0.1 - - [19/Oct/2021 08:31:47] "GET / HTTP/1.1" 200 -
127.0.0.1 - - [19/Oct/2021 08:31:48] "GET /_dash-layout HTTP/1.1" 200 -
127.0.0.1 - - [19/Oct/2021 08:31:48] "GET /_dash-dependencies HTTP/1.1" 200 -
127.0.0.1 - - [19/Oct/2021 08:31:48] "GET /_dash-component-suites/dash/dcc/async-dropdown.js HTTP/1.1" 200 -
127.0.0.1 - - [19/Oct/2021 08:31:48] "GET /_dash-component-suites/dash/dcc/async-graph.js HTTP/1.1" 200 -
127.0.0.1 - - [19/Oct/2021 08:31:48] "GET /_dash-component-suites/dash/dcc/async-plotlyjs.js HTTP/1.1" 200 -
127.0.0.1 - - [19/Oct/2021 08:31:48] "POST /_dash-update-component HTTP/1.1" 200 -
```

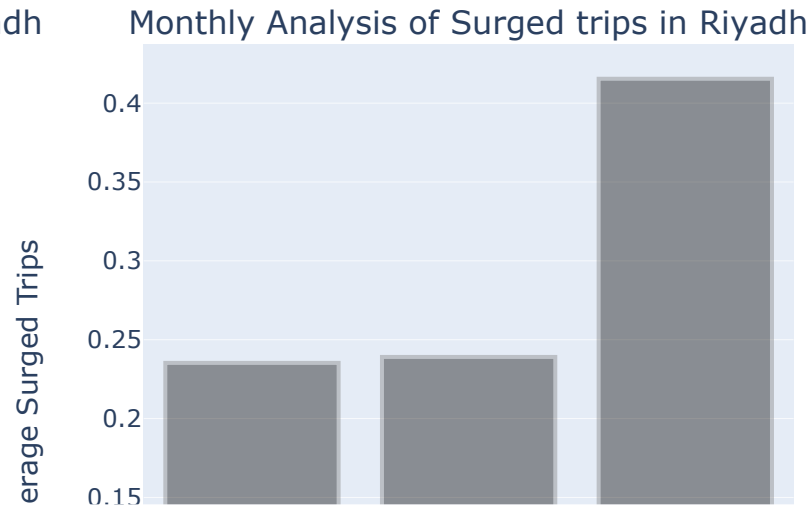
e. Monthly analysis of completed and surged trips

Monthly analysis of the completed trips and surged trips is done to give a comparison of the trips based on months. The data has three months: May, June and July. However, the code has been integrated to include all the months to include the data on an ongoing basis.

Out[94]:

	Month	completed_trip	surged_trip
0	May	0.992334	0.235313
1	Jun	0.993353	0.239085
2	Jul	0.996779	0.415459





Interactive tool - Monthly analysis of completed and surged trips

To see the interactive visualization, please click on the below last link:

```
Dash is running on http://127.0.0.1:8050/ (http://127.0.0.1:8050/)

Dash is running on http://127.0.0.1:8050/ (http://127.0.0.1:8050/)

Dash is running on http://127.0.0.1:8050/ (http://127.0.0.1:8050/)

Dash is running on http://127.0.0.1:8050/ (http://127.0.0.1:8050/)

Dash is running on http://127.0.0.1:8050/ (http://127.0.0.1:8050/)

Dash is running on http://127.0.0.1:8050/ (http://127.0.0.1:8050/)

 * Serving Flask app "__main__" (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production deployment.
   Use a production WSGI server instead.
 * Debug mode: off

 * Running on http://127.0.0.1:8050/ (http://127.0.0.1:8050/) (Press CTRL+C to quit)
127.0.0.1 - - [19/Oct/2021 08:34:10] "GET / HTTP/1.1" 200 -
```

f. Analysis based on demand and supply disparity - Vehicle type

a. **Findings**: Based on table, the gap in number of Black Car requested vs the number of Black Car being the vehicle type is more than 50%. Requested #: 8692 and Vehicle Type #: 94. The demand is not met here and instead UberX is sent for the request of Black Car. From Second table (refer the table in next analysis), it is seen that Average fare/trip of a Black Car is 3 times more than the average fare/trip of UberX

b. **Assumption**: If the rider is requesting an Uber black but getting an UberX then then the assumption is being made that rider is being charged for UberX

c. **Analysis**: The revenue can be increased by supplying Black Car, whenever requested, at least by 50% (4.5k), leading to additional revenue of $38k (22% additional revenue). Due to the disparity in vehicle type, not only Uber is bearing the losses but the rider is also dissatisfied, leading to decrease in Uber's brand value.

d. **Recommendation**: Uber could focus on increasing the number of Black Cars to make sure that whenever there is a demand for Black Car, it is met.

Library to install to run the below interactive table: pivottablejs

PivotTable.js is a Javascript Pivot Table and Pivot Chart library with drag'n'drop interactivity, and it can now be used with Jupyter/IPython Notebook via the pivottablejs module.

```
Requirement already satisfied: pivottablejs in c:\programdata\anaconda3\lib\site-packages (0.9.0)
```

Out[103]:

[pop out]

Table | trip_status | city_id | trip_fare_usd

Sum | city_id | vehicle_type

request_type

request_type	vehicle_type	Black Car	UberVIP	UberX	null	uberXL	Totals
					0.00		**0.00**
Black		94.00	17.00	8,501.00		80.00	**8,692.00**
UberX		577.00	241.00	27,272.00		1,514.00	**29,604.00**
	Totals	**671.00**	**258.00**	**35,773.00**	**0.00**	**1,594.00**	**38,296.00**

g. Analysis of Average Revenue generated based on Type of Vehicle

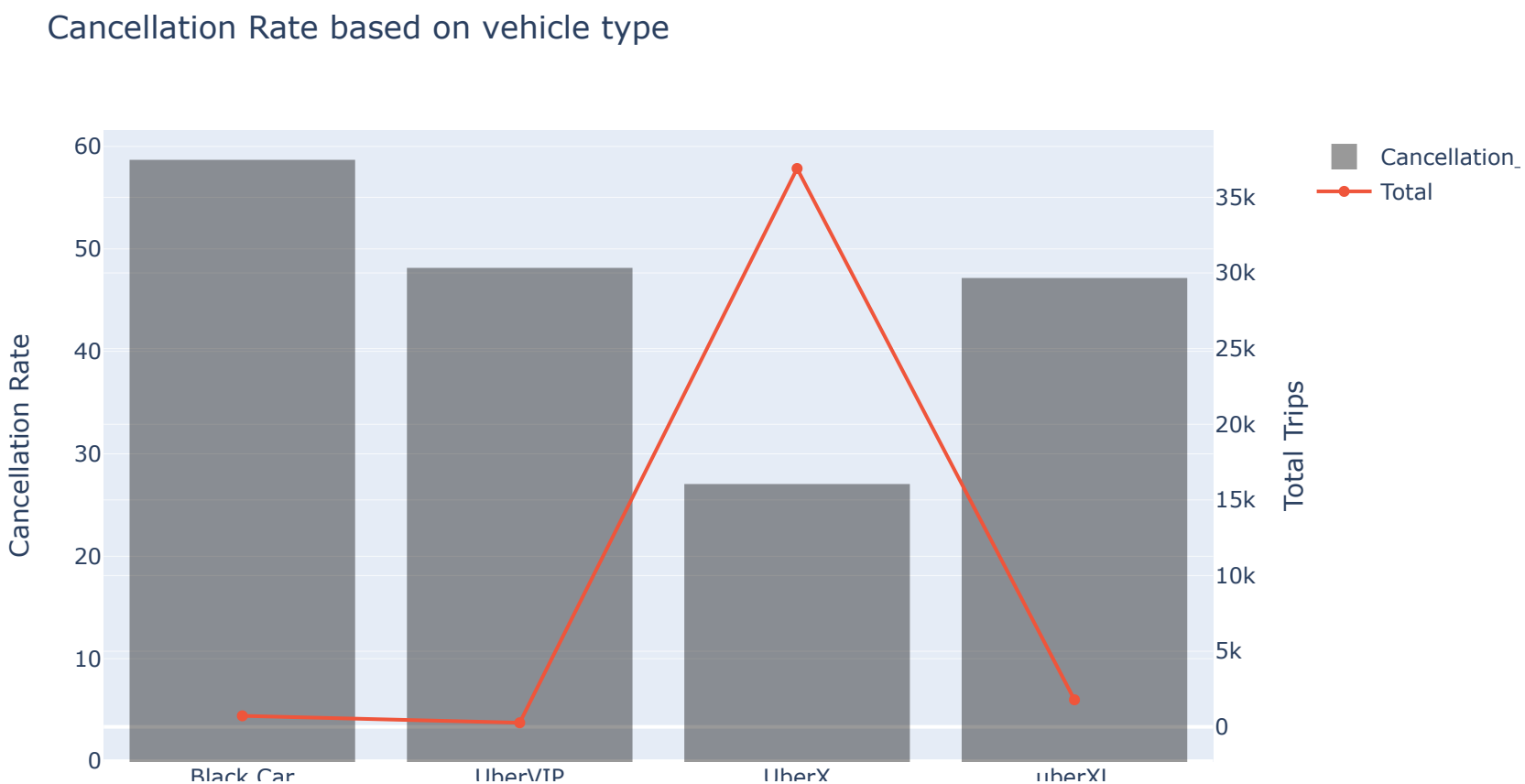
Out[110]:

	vehicle_type	trip_fare_usd	Total trips	Avg revenue per vehicle type
0	Black Car	4427.328621	302	14.660029
1	UberVIP	762.200990	141	5.405681
2	UberX	160754.736516	26987	5.956747
3	uberXL	10562.646521	949	11.130291

h. Analysis of cancellation rate vs type of vehicle

Out[113]:

	vehicle_type	Total_Cancelled_trip	Total	Cancellation_rate_vehicle_type
0	Black Car	426	728	58.516484
1	UberVIP	130	271	47.970480
2	UberX	9917	36904	26.872426
3	uberXL	841	1790	46.983240



a. **Finding**: Despite being the highest requested vehicle type, UberX is cancelled very less number of times compared to Black, VIP and XL.

b. **Hypothesis**: There could be a couple of reasons for the other vehicle types to get cancelled frequently:

1. Either the black cars are less in number, and multiple requests are going to them simultaneously, and the driver has to cancel some of the requests.
2. The black cars are not available in the near location and is taking longer wait times, and the rider itself is cancelling the ride.

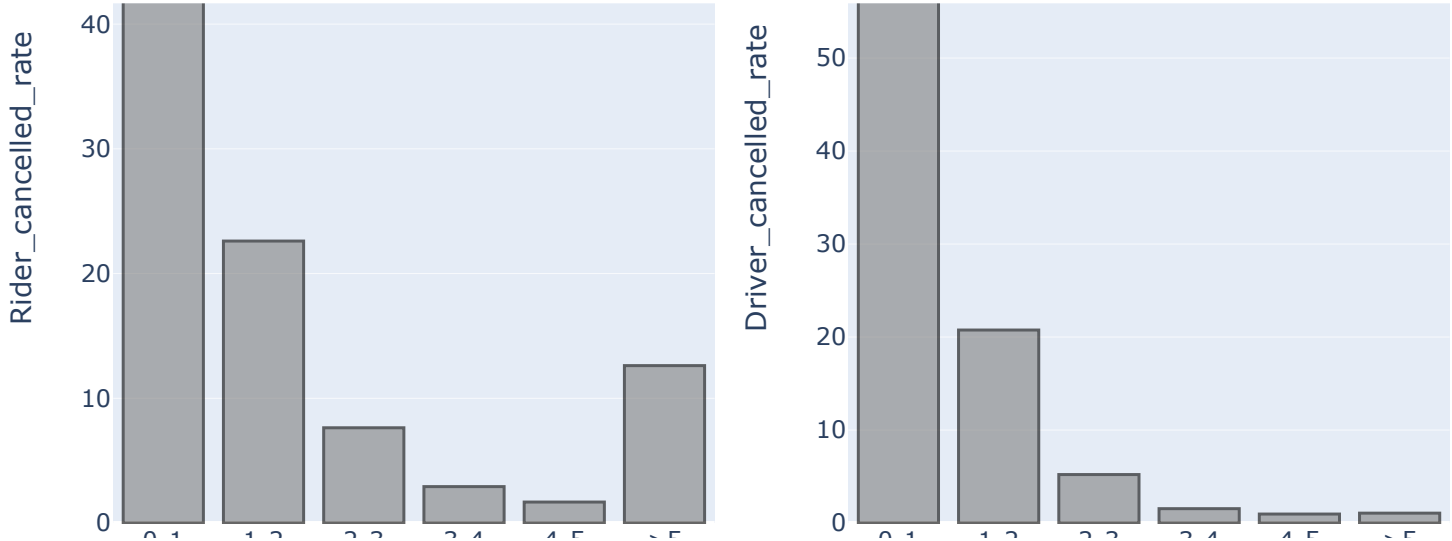
c. **Recommendation**:

1. Addition to the fleet could minimize the cancellation of rides.
2. Better Allocation of other vehicle types with respect to location will lead to reduction in wait time for the rider

Secondary Analysis

Analysis of Cancellation Rate based on Rider and Driver

Out[125]:
```
Distance_pick_up_window
0-1    5345
1-2    2300
2-3     777
3-4     297
4-5     171
>5     1284
Name: trip_status, dtype: int64
```



a. **Finding**: Here, the rides are getting cancelled within 0-1 mile distance to pickup window by both the riders and drivers.

b. **Analysis**: Various hypothesis can be built based on the time taken by the rider to cancel after a ride is being booked, if more data was provided.

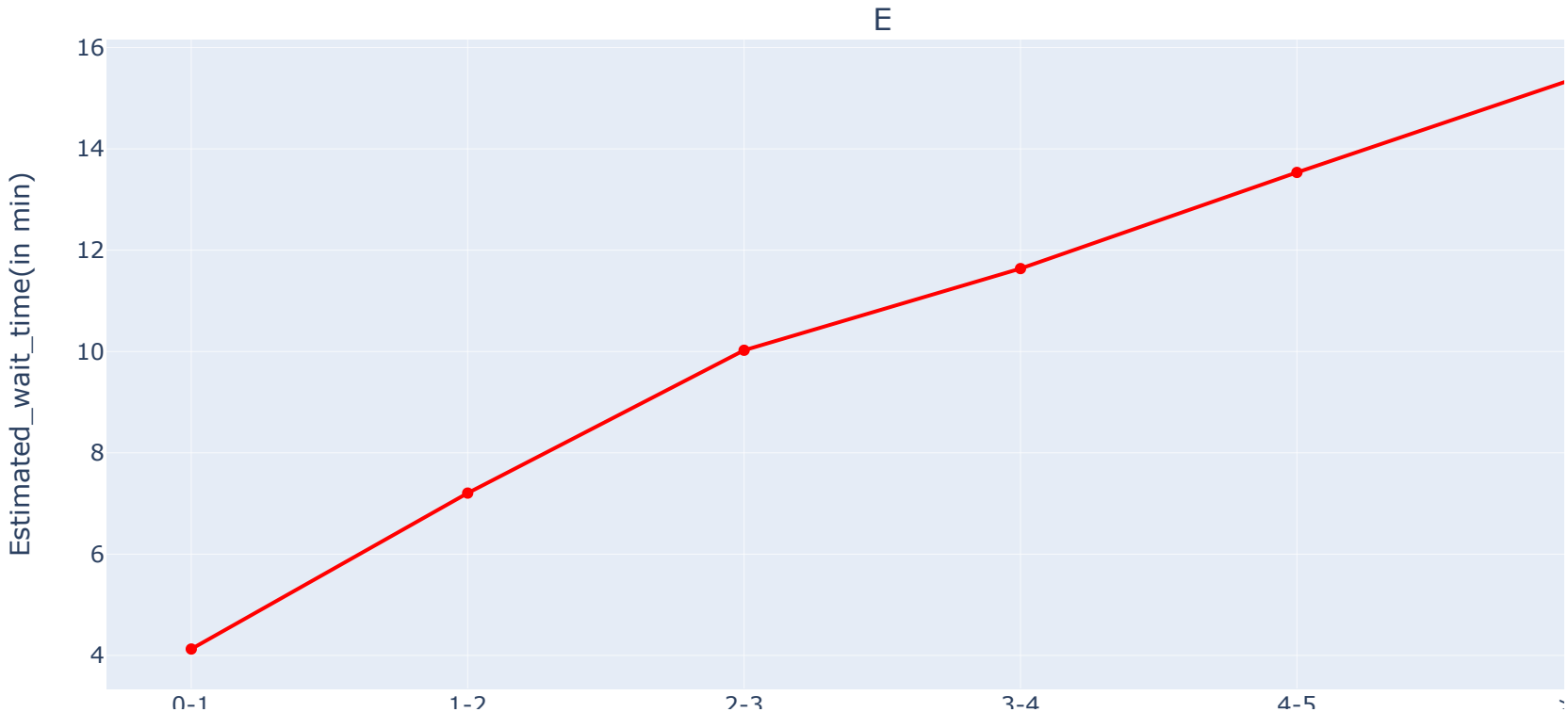
c. **Recommendation** could be to provide a feedback form with some predefined options like:

1. Booked accidentally
2. Cancelled accidentally
3. Changed my mind
4. Got better deals in other ride-hailing apps/services.. etc

Facts based on pickup time vs estimate time to pickup

Out[134]:

	Distance_pick_up_window	esttime_to_pickup
0	0-1	4.126107
1	1-2	7.204435
2	2-3	10.024603
3	3-4	11.637093
4	4-5	13.535185
5	>5	15.377690



Analysis based on Mode of payment

Out[141]:

	paid_cash	Total	Total_pct
0	No cash	6424	16.06
1	cash	33576	83.94

Riders in Riyadh are paying in either cash or via apps, credit/ debit cards, coupons.

a. **Findings**: Based on data, 84% of the rides are paid by riders in cash

b. **Recommendation**: Uber can build better relations with Riyadh, as Saudi Arabia is steadily moving towards building a Cashless Society by 2030, by encouraging riders to pay in cash via promotions/ offers on Uber. This can be achieved by establishing relations with companies like MasterCard, Visa, PayPal, or any local financial services.