Data Product Manager Nanodegree

Applying Data Science to Product Management

Midterm Project: Developing an MVP Launch Strategy for a Flying Taxi Service

Welcome to your first week at Flyber


Aybar

In this project, you will apply the skills acquired in this course to create the MVP launch strategy for the first flying car taxi service, Flyber, in one of the most congested cities in America -- New York City.

You are responsible for bringing the first flying car taxi service to market by analyzing data and building a product proposal.

You will need to use the SQL workspace provided in the Classroom, and Tableau Public, in order to successfully complete the project.

You'll present your answers, findings, and insights in the Answer Slides found in this deck. Feel free to include any additional slides, if needed.

Section 1: Data Exploration

Back to the basics of product management, identify your customer and their pain points:

- What are taxis used for?
- What are the characteristics of the users that leverage them?
- What are existing pain points with taxis?
- What are the existing pain points with digital ride-sharing services?

Answer Slide

- Taxis are used for mobility and take users from point a to point b.
- Usually taxi users are people who want comfort and mobility, people who accept to pay more than public transport to make it easier to get around the city.
- Taxi users usually face great competition to obtain taxis at peak times. In addition, of course, concern for safety during the trip.
- Applications do not work in all locations, usually inland cities do not have digital travel sharing services.

What user improvements do you hypothesize a flying taxi service would have over the existing state of taxis today?

What market improvements do you hypothesize a flying taxi service would have the existing taxi service industry & physical road infrastructure today?

Answer Slide

A flying taxi would have several advantages, one of which is to get rid of the heavy traffic in New York, for sure users will have a faster time getting around, it will also make life easier for those users who live a little further away from urban centers.

As there will be less taxis on the street, asphalt wear will be less, it will also open up several new job opportunities for the logistics involved in an air taxi in New York.

Upload this dataset into Tableau Online.

Ensure the fields are parsed correctly; field headers are included in the first row of the CSV.

Let’s begin exploration!

Acquire a high-level understanding of the granularity and scope of the dataset, to inform the basis for your analyses:

- How many records are in the dataset
- What does each record represent?
- What is the primary key?
- What date range is your dataset bound to?
- What are the geographical bounds of this dataset? Is it limited to Manhattan, or is Brooklyn, Queens, Staten Island, the Bronx, and New Jersey included? Where are most of the data points centralized at? Are there outliers?

Answer Slide

The dataset has 8 columns, seeing the count attribute of each column we can say that the dataset has 145,864 records.

```
In [3]: df.describe()
```

Out[3]:

	vendor_id	passenger_count	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	duration	distance
count	145864.000000	145864.000000	145864.000000	145864.000000	145864.000000	145864.000000	145864.000000	145855.000000
mean	1.534217	1.660060	-73.973667	40.750939	-73.973571	40.751842	945.583756	3.431454
std	0.498830	1.308973	0.136206	0.032790	0.136673	0.038286	3106.087860	4.369522
min	1.000000	0.000000	-121.933342	36.118538	-121.933304	36.118538	1.000000	0.000000
25%	1.000000	1.000000	-73.991943	40.737381	-73.991302	40.735985	397.000000	1.231608
50%	2.000000	1.000000	-73.981735	40.754086	-73.979767	40.754635	661.000000	2.091339
75%	2.000000	2.000000	-73.967270	40.768452	-73.963142	40.770008	1072.000000	3.871287
max	2.000000	6.000000	-61.335529	41.301289	-61.335529	41.427902	86366.000000	578.842818

Answer Slide

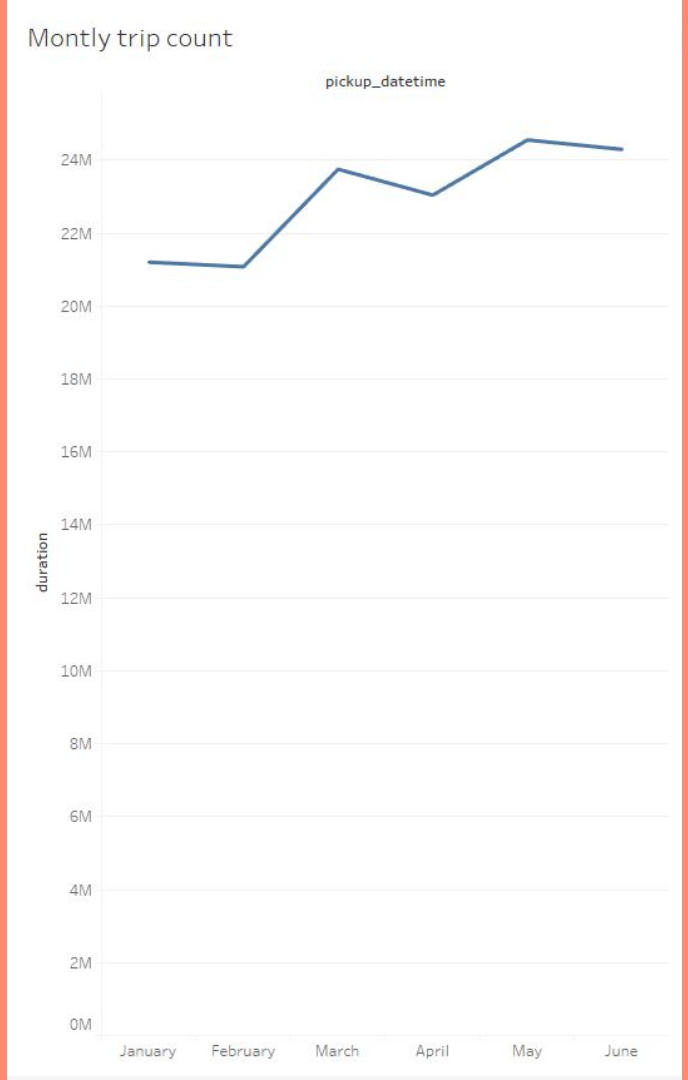
Each dataset line consists of a taxi ride. I searched Kaggle for the description of the dataset columns and put it below.

Data fields

id - a unique identifier for each trip
vendor_id - a code indicating the provider associated with the trip record
pickup_datetime - date and time when the meter was engaged
dropoff_datetime - date and time when the meter was disengaged
passenger_count - the number of passengers in the vehicle (driver entered value)
pickup_longitude - the longitude where the meter was engaged
pickup_latitude - the latitude where the meter was engaged
dropoff_longitude - the longitude where the meter was disengaged
dropoff_latitude - the latitude where the meter was disengaged
store_and_fwd_flag - This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip
duration - duration of the trip in seconds

Answer Slide

This view shows the distribution of the number of trips per month.



Answer Slide

The preview on the side shows a report on the pickup distribution by Boroughs in New York, which shows that the large number of trips is initiated in the Manhattan region.

You notice that the dataset does not contain explicit data points out-of the-box, we'll need to enrich the dataset with relevant fields:

- You notice that ride price is not included, but figure it could be derived. Based on information about New York taxi prices gleaned from the internet, create a calculated field called `price` using the `duration`, `distance`, and `passenger count` fields.
- You hypothesize your target users will be those who take a relatively longer time getting to a destination that is relatively close, due to heavy traffic conditions and/or limitations to physical road infrastructure. To be able to analyze where this is happening, you will need to create a calculated field called `distance-to-duration ratio`.

Answer Slide

duration	distance	preco	distance-to-duration_ratio
455	1.498521	10.055351	0.329345
1274	3.806139	22.676166	0.298755
486	2.505926	12.843884	0.515623
1479	4.564593	26.289432	0.308627
1450	5.477047	28.339483	0.377727

To calculate price i used Python:

df['preco'] = (2.5 +(1.56*df.distance*1.61)+(df.duration /3600)*30)

So, to calculate the distance-to-duration_ratio i used:

df['distance-to-duration_ratio'] = df.distance / df.duration * 100
When I didn't multiply by 100, the instructor reported that the data was out of the standard.

It´s correct?

Let’s understand the scope and distribution various dimensions within the dataset. Calculate the **average**, **median**, and the **first & second standard deviation of the mean** for the following measures:

- duration
- distance
- passenger counts
- duration-to-distance ratio
- price

Answer Slide

I used the describe function to analyze the duration, distance, passenger_count, price and distance-to-duration_ratio columns

```
In [42]: df_describe.describe()
```

Out[42]:

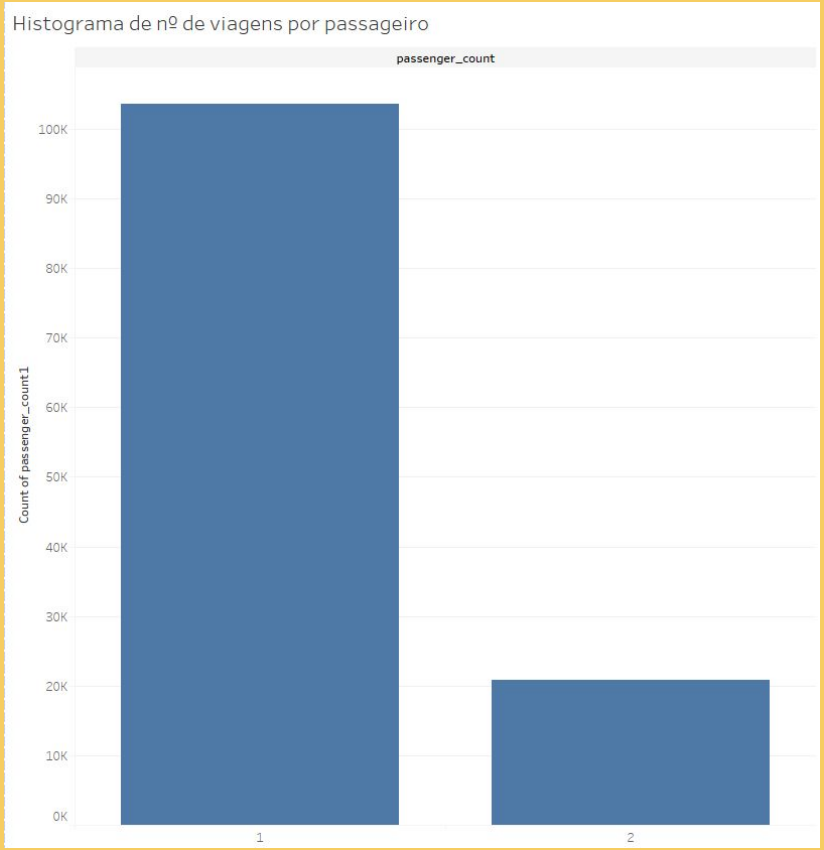
	duration	distance	passenger_count	price	distance-to-duration_ratio
count	145855.000000	145855.000000	145855.000000	145855.000000	145855.000000
mean	945.611059	3.431454	1.660039	18.998532	0.401080
std	3106.180228	4.369522	1.308974	29.546786	0.559415
min	1.000000	0.000000	0.000000	2.508333	0.000000
25%	397.000000	1.231608	1.000000	9.222771	0.253388
50%	661.000000	2.091339	1.000000	13.541227	0.355549
75%	1072.000000	3.871287	2.000000	21.215264	0.495923
max	86366.000000	578.842818	6.000000	1458.846620	191.037233

Flying cars may have to have to be a lower weight for efficiency & take-off. Or you may just decide to leverage mini-copters for your initial MVP.

Create a histogram that visualizes the number of total rides grouped by passenger counts to analyze the potential market volume of low passenger pickups (1-2 passengers).

Answer Slide

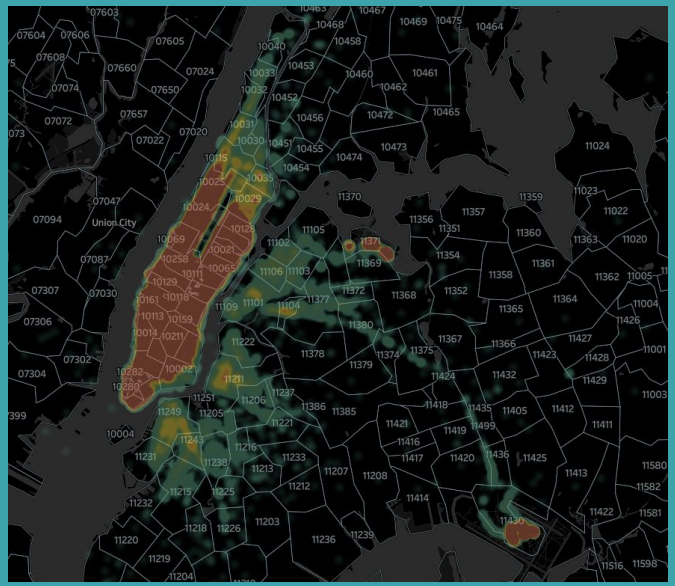
In the visualization we have the histogram that compares the passenger count for trips with one or two passengers.



For the initial MVP launch (& most likely GA), we have a finite amount of monetary resources to build Flyber pick-up / drop-off nodes. We'll need to be strategic on where we'll place them:

- Which neighborhoods/zip codes tends to experience a relatively higher density of pick-ups?
- Which neighborhoods/zip codes tends to experience a relatively higher density of drop-offs?
- Which neighborhoods/zip codes tends to have the highest duration-to-distance ratios, based on pick-up?
- Which neighborhoods/zip codes tends to have the highest duration-to-distance ratios, based on drop-off?
- For any of the neighborhoods identified, are there any potential areas within the neighborhood that are optimal for flying taxi pick-up / drop-off? What makes them suitable?

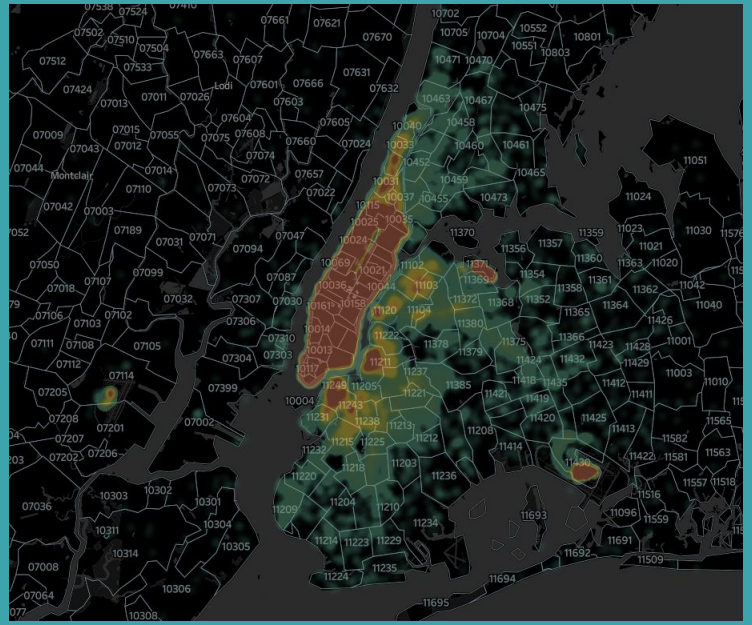
Answer Slide - Pickup analysis



The two pickup views show that the largest number of completed trips takes place in the areas: 10280 - 10282 - 10002 - 10211 - 10113 - 10161 - 10118 - 10129 - 10111 - 10258 - 10065 - 10021 - 10069 - 10024 - 10128 - 10025 - 10029

We can also see some pickup locations farther from downtown Manhattan, at locations 11430 e 11371.

Answer Slide - Dropoff analysis



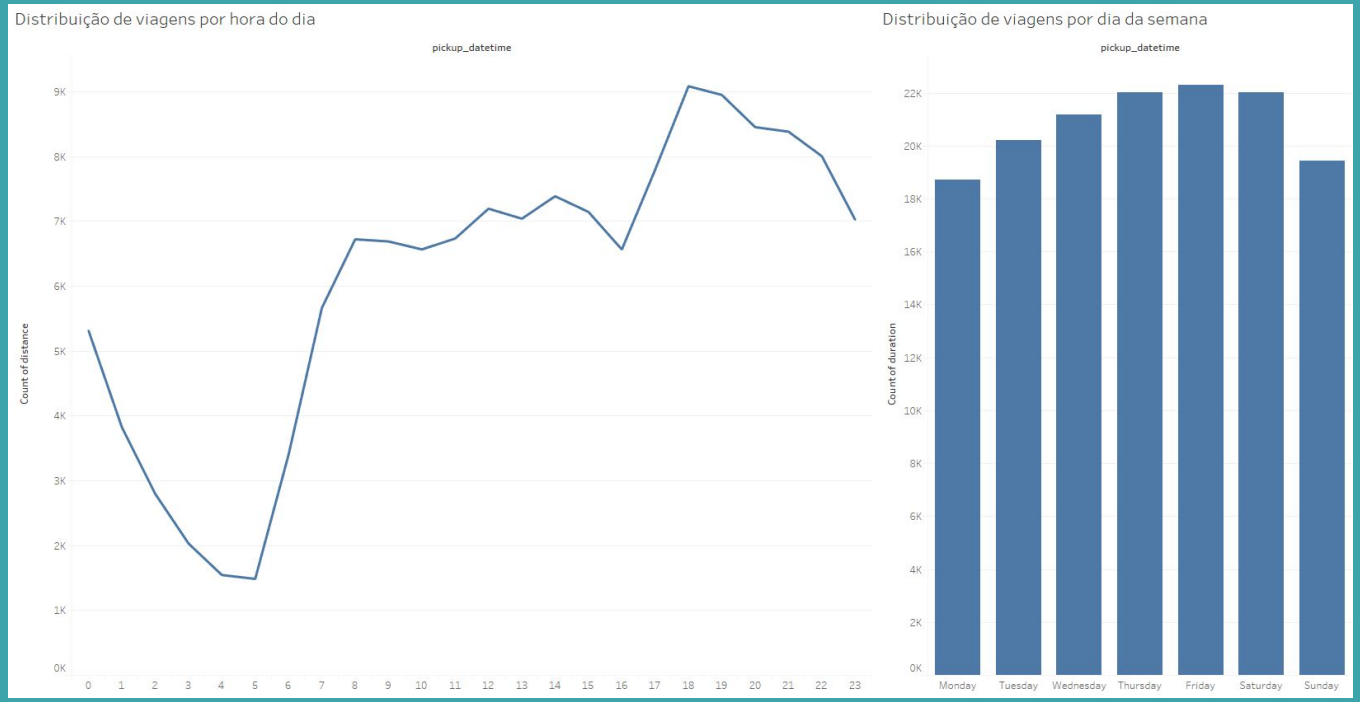
The two dropoff views show that the largest number of completed trips takes place in the areas: 10117 - 10013 - 10014 - 10161 - 10158 - 10036 - 10021 - 10024 - 10025 - 10035 - 10115 - 12211

We can also see some dropoff locations farther from downtown Manhattan, at locations 11430 and 07114.

It may not make operational sense to have the service running 24/7, for now.

- What times throughout the day experience relatively higher volumes of ride pick-ups?
- What days throughout the week experience relatively higher volumes of ride pick-ups?
- Pinpoint any periods throughout the year that experience trend fluctuation or seasonality around ride pick-up volumes. This will help us in our post-launch analyses to determine if any spikes or dips were influenced by seasonality or through actual feature adoption/regression.

Answer Slide



Analyzing the pickup data by the time of day, we can clearly see that the peak is around 7 pm.

While analyzing the travel count by day of the week, we can see that Thursday and Friday are the days of the week that have the most trips.

You and the user research team ran a quantitative survey on existing taxi and/or rideshare users in New York City to determine sentiment around potentially using a flying taxi service.

Dive into the survey results dataset in order to extract insights from explicit feedback.

Upload this dataset into Tableau Online or a SQL database (the classroom contains a workspace with the data for you as well).

Ensure the fields are parsed correctly, field headers are included in the first row of the CSV.

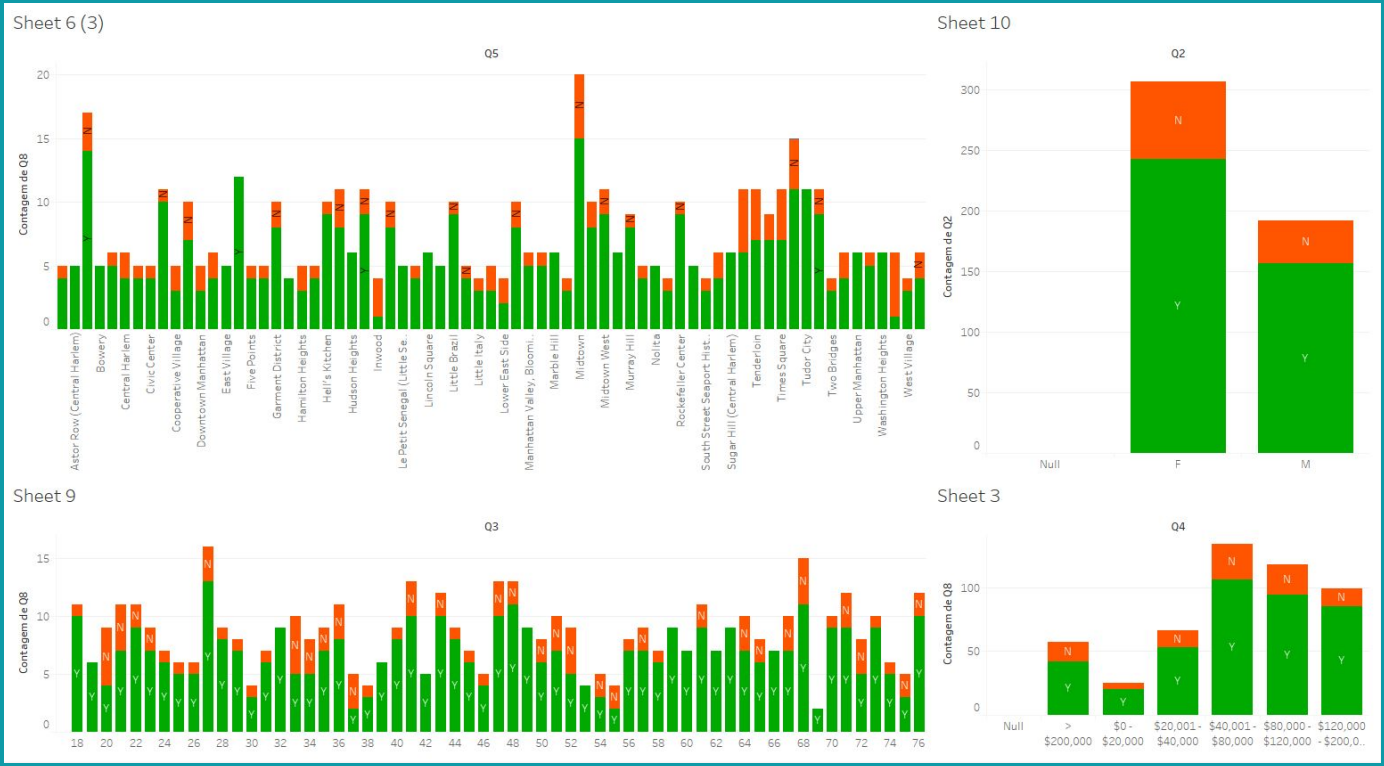
Question schema:

Q1 - What is your email?
Q2 - What gender do you identify as?
Q3 - What is your age?
Q4 - What is your annual income? (income bands)
Q5 - What neighborhood do you reside in?
Q6 - Do you currently use taxis? (Y/N)
Q7 - Do you currently use ridesharing services? (Y/N)
Q8 - Would you use a flying taxi service, if such a concept existed? (Y/N)
Q9 - If yes to Q8, how much would you be willing to pay per mile for such a service? (USD)
Q10 - If no to Q8, what is the reason?

To inform our future product marketing efforts, we'll want to extract the following:

- Is there an inclination of better Flyber adoption based on gender, age, income level, or neighborhood of residence?
- What is the distribution of potential price per mile based on gender, age, income level, and neighborhood of residence?
- What is the different personas/segments of negative sentiment towards not using a flying taxi car service?

Answer Slide

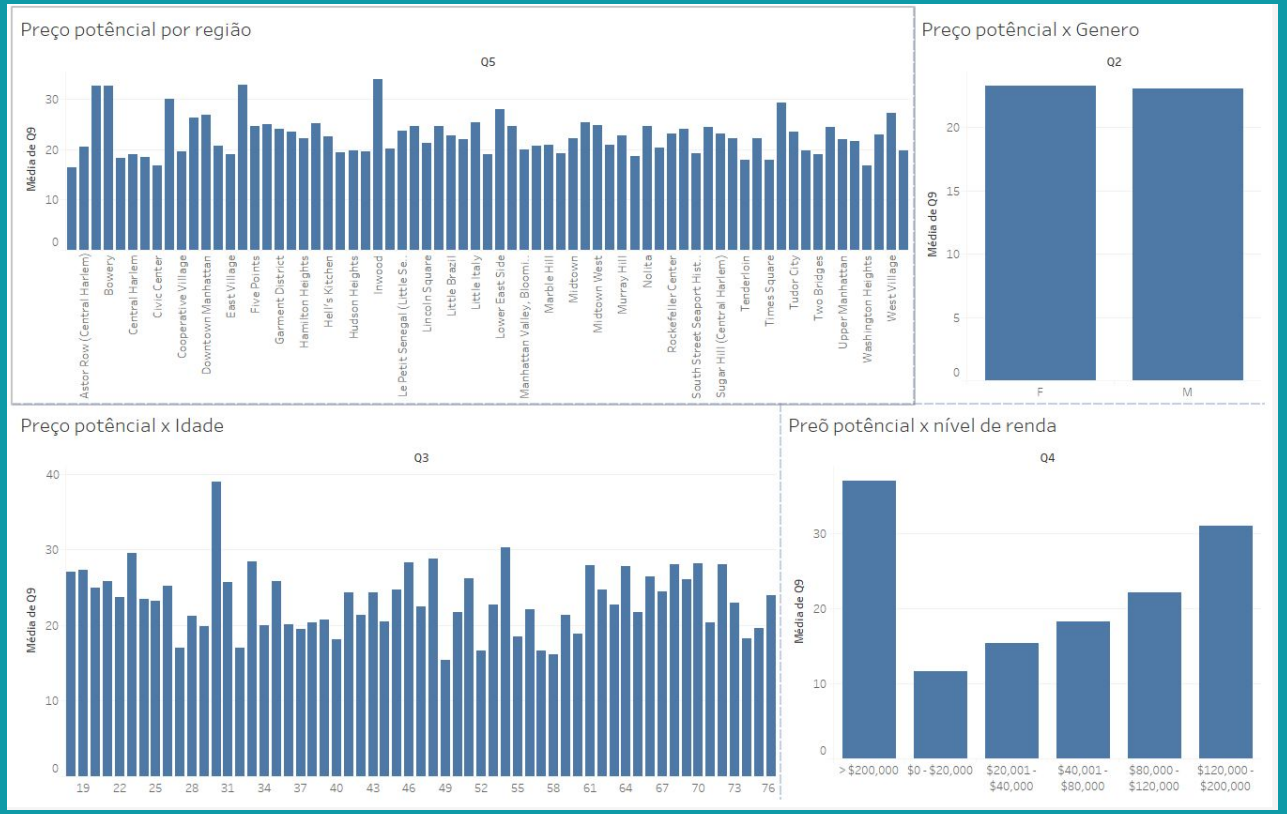


In this view, I crossed the data of people who answered yes or no to question 8 of the questionnaire.

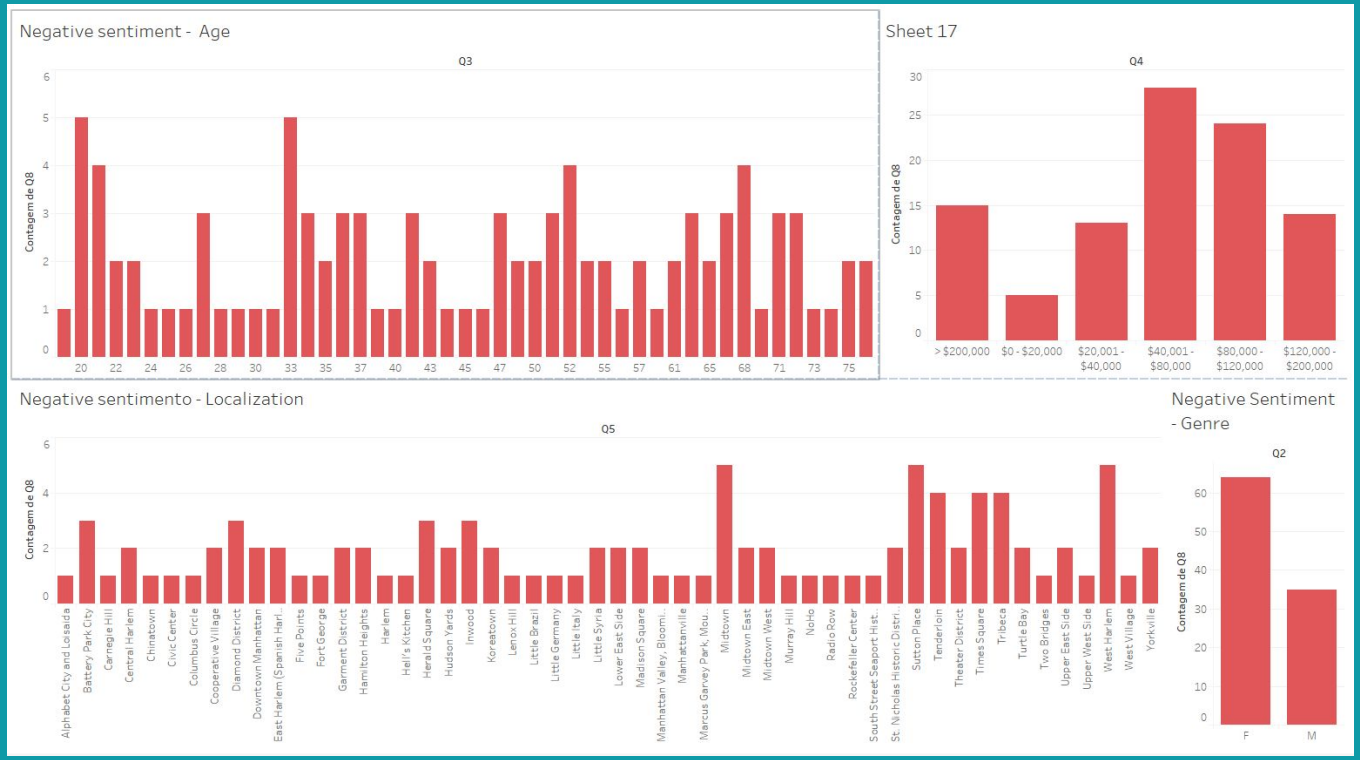
I crossed the data based on the region of the pickups, age, sex, and how much I would pay for the service, used green and red to make the discrepancy between yes and no more visually clear.

Answer Slide

In this view, I crossed the price counting data per mile based on the region, sex, age and price I would pay for the flying taxi service;



Answer Slide



Hooray! End of Section 1.

You will complete Section 2 at the end of this course.

Please submit this file for review for Section 1.

Section 2: Proposal Synthesis

Identify a product objective for Flyber's launch. Your product objective will guide your KPIs, so identify what Flyber should optimize for. Your objective should be centered around one the following focus areas:

- User Acquisition
- User Engagement
- User Retention
- Profitability

Explain your reasoning. Include both why you feel your focus area is more relevant than the others for Flyber at this time of the product development cycle.

Answer Slide

(Fill out your answer here)

Formulate 3-5 Key Performance Indicators (KPIs), to measure if the product is heading towards the right direction based on your objective

Answer Slide

(Fill out your answer here)

Create hypotheses around what thresholds your KPIs would need to hit in order to determine success

Answer Slide

(Fill out your answer here)

As the product manager, you make decisions based on the insights you extract, we'll need to know the feature set we'll include in the MVP to measure viability, while keeping operational expenditure under control:

- What times/days of operation should the service run for?
- How many pick-up / drop-off nodes should we have?
- Where should the nodes be located?
- Should we initially use copters or homegrown hardware?
- Should the pricing be fixed or dynamic? At what rates?

Answer Slide

(Fill out your answer here)

Determine the MVP sample size & time period allotted estimated to come to a conclusion on your hypotheses.

Answer Slide

(Fill out your answer here)

Create an instrumentation plan for the events you need collected and logged, in order to be able to physically measure your KPIs.

Answer Slide

(Fill out your answer here)

Create a qualitative feedback survey questions for users after their ride, to further understand and optimize the product for future iterations.

Answer Slide

(Fill out your answer here)

Summarize everything you have learned into your final proposal

- Identify the target population. Why did you select that target population? What are their pain points?
- Create a product proposal containing claim, evidence, estimated impact, and risks
- Claims should be backed by quantitative evidence, impact should assess market needs/benefits
- Risks involve any known unknowns that we'll still need to monitor post-launch
- State cross-functional stakeholder teams that will need to be involved

Answer Slide

(Fill out your answer here)

Answer Slide

(Fill out your answer here)

Answer Slide

(Fill out your answer here)