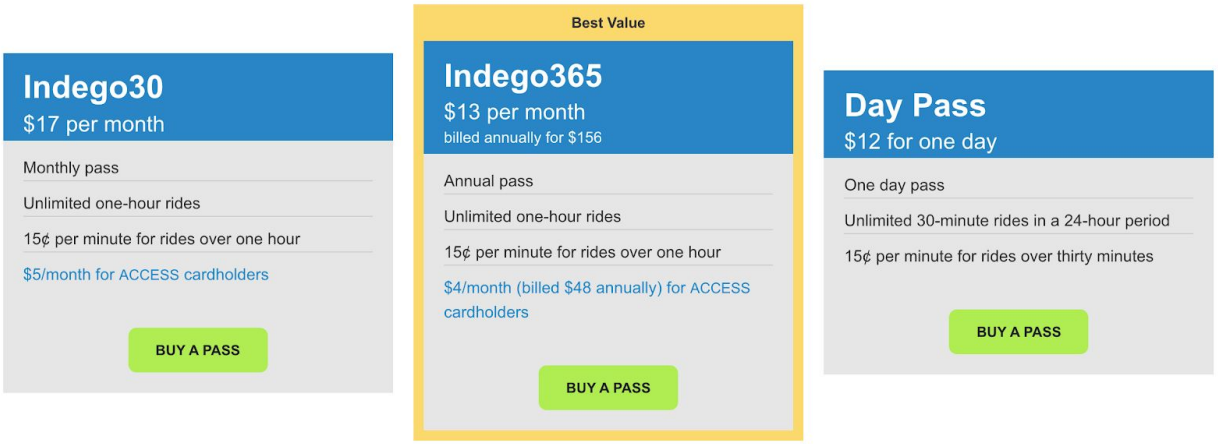
Gregory Walsh
Project 3 - Indego Bike Data
Fall 2019 Intro to Data Science

Indego is a bike sharing company for the Philadelphia region that provides a healthier and environmentally friendly alternative to Uber or Lyft; providing traditional and electric bikes as a means of transportation.

In an effort to improve their services and better serve their customers, they provide something they call "Trip Data". The Indego Trip Data is open source data on the bike rides in Philadelphia. They encourage people to download the data and discover the hidden gems there might be within the data. When I first looked at the data I was interested in how their price plans worked.



The most popular price plans are the “Indego 30”, “Indego 365”, and “Day Pass/Walk Up”. The questions that I wanted to answer were what plan is being utilized the most and is the “Indego 365” truly the “best value” based on the Trip Data?

Disclaimer: I am in no way affiliated with Indego, I have no stake in how the company operates or how much money they make. This was an independent analysis for my Masters in Data Science and is my own observation and should be criticized as such.

In order to answer my questions, I needed to first scrape and clean the csv files that sit on Indego’s website. For this project I used BASH to collect and clean the data. From there I used Python to explore and visualize the data.

Here is the Bash script:

```
#! /bin/bash
curl https://www.rideindego.com/about/data/ > bikedata
grep -E -o 'http.*\.zip' bikedata > urls
```

```
echo on
readarray -t arr < urls
for i in "${arr[@]}"
do
echo $i
done
### Replace wd with your own
indegodata=$(ls "/Users/gregwalsh/Github/DataScience
/DSSA-5001-Intro-To-DSSA/week_10")
cat $indegodata > indegocombined.csv
for file in $indegodata
do
    continue
    ### Clean the Data of " and remove the col's of every file other than
one
    if [ ${file: -4} == ".csv" ]
    then
        echo $file
        sed -i '' 's/\"//g' $file
        head -n 1 $file
    fi
    ### Take one of the more recent datasets and set that as the header in
the new csv to be made.
    ### From there remove the first line (Columns) in each other csv
    if [ $file == "indego-2019-q2.csv" ]
    then
        echo "Found"
        cat $file > indegocombined.csv
    else
        sed -i '' 1d $file
    fi
    ### Finally cat each csv and append it to a new file
    cat $file > indegocombined.csv
done
```

Once the Bash script is done there was a csv called indegocombined.csv with all the ride data from 2015 - 2019.

Before I moved onto Python, I looked at the data columns.

```
➜ head -n 1 indegocombined.csv
```

```
trip_id,duration,start_time,end_time,start_station,start_lat,start_lon,end_station,end_lat,end_lon,bike_id,plan_duration,trip_route_category,passholder_type,bike_type
```

The data used is described on Indego's website here.

- **trip_id:** Locally unique integer that identifies the trip
- **duration:** Length of trip in minutes
- **start_time:** The date/time when the trip began, presented in ISO 8601 format in local time
- **end_time:** The date/time when the trip ended, presented in ISO 8601 format in local time
- **start_station:** The station ID where the trip originated (for station name and more information on each station see the *Station Table*)
- **start_lat:** The latitude of the station where the trip originated
- **start_lon:** The longitude of the station where the trip originated
- **end_station:** The station ID where the trip terminated (for station name and more information on each station see the *Station Table*)
- **end_lat:** The latitude of the station where the trip terminated
- **end_lon:** The longitude of the station where the trip terminated
- **bike_id:** Locally unique integer that identifies the bike
- **plan_duration:** The number of days that the plan the passholder is using entitles them to ride; 0 is used for a single ride plan (Walk-up)
- **trip_route_category:** "Round Trip" for trips starting and ending at the same station or "One Way" for all other trips
- **passholder_type:** The name of the passholder's plan
- **bike_type:** The kind of bike used on the trip, including standard pedal-powered bikes or electric assist bikes

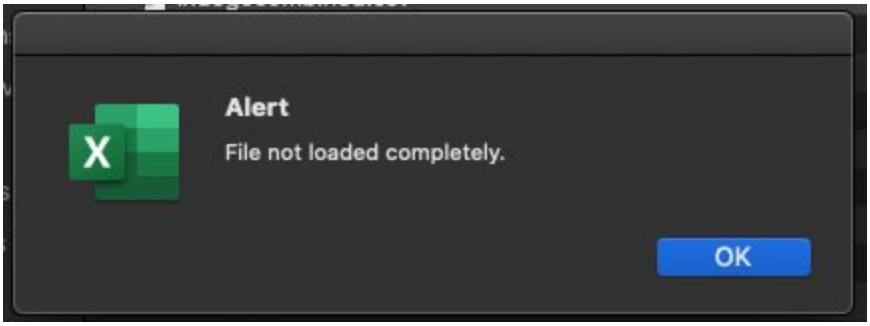
The columns I was Interested in are the **start_time** or **end_time** to get the year and the **passholder_type** looking specifically for the three plans sold above.

It is important to look at the amount of data you are working with to see what kind of compute you will need to process the data manipulation.

```
➜ wc -l indegocombined.csv
 3,123,637 indegocombined.csv
```

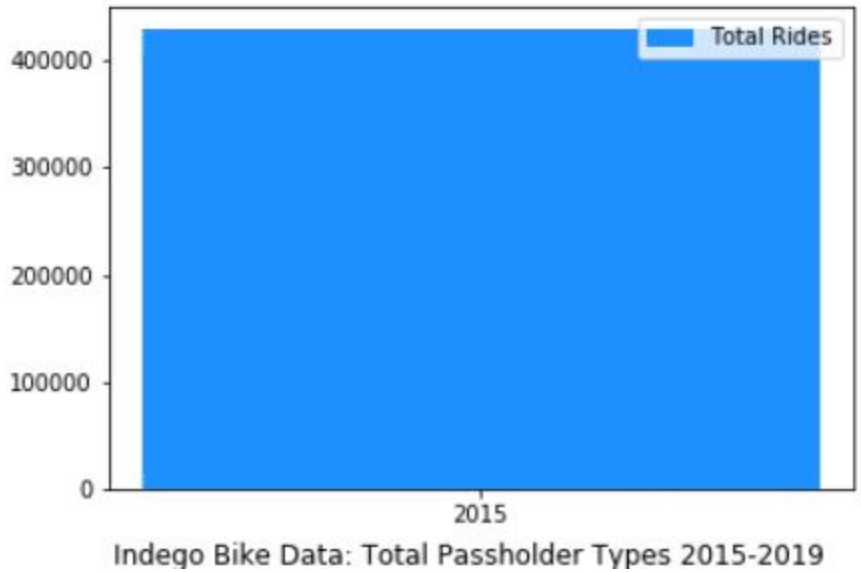
At this point, I was working with over three million rows of data. Working with this mass amount of data and the formatting of it are some of the many reasons learning Bash is helpful when cleaning and combining data.

Side Note: Try loading the csv in Excel and you will get this error showing Excel hit its max row limit:



Next I headed into a Python Environment to explore and visualize the data.

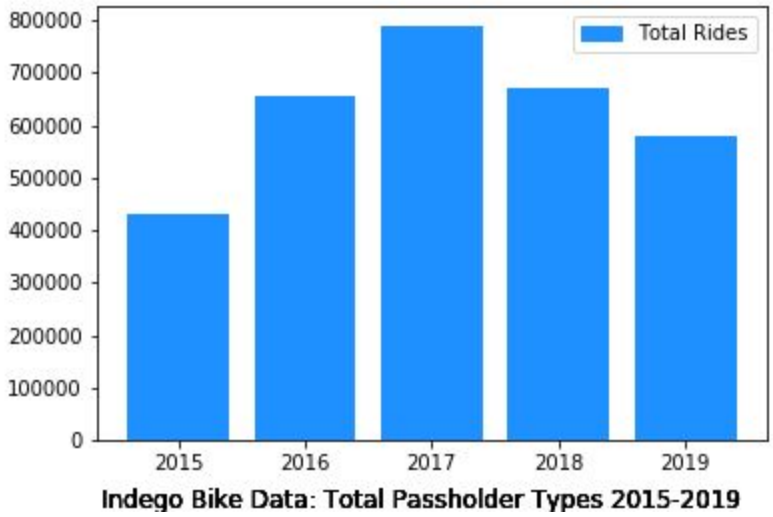
As part of this process, I used Jupyter Notebook inside of VSCode and Anaconda as my Python Package Manager. The script used for this project will be in the link below for anyone that would like to explore the data for themselves.



The script will create the following visualization:

The visualization above shows the Total Passholder Types per year.

Below, starting with blue, the total rides are plotted per year:



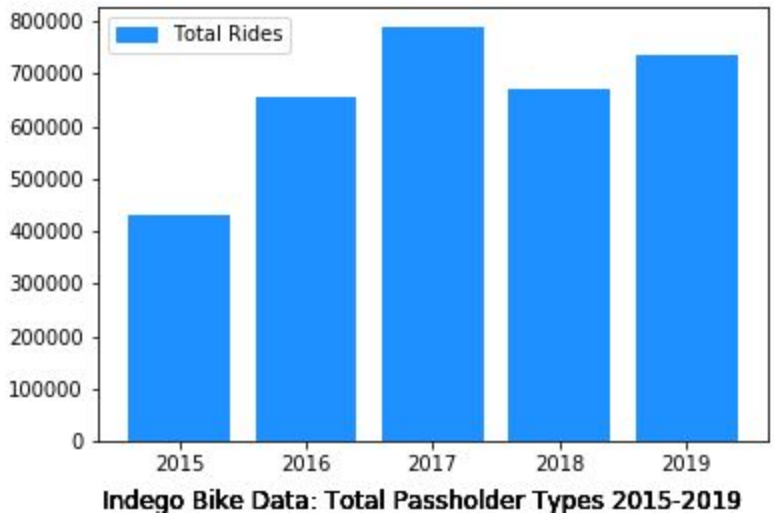
Interestingly, at this point in my analysis, it would seem that Indego is on a downward trend for total rides according to the trip data. It is important to note that this trend **should not** be correlated with Indego's total business revenue since we are looking at a partial dataset. Since the data for 2019 only goes until Q3 (July 2019– September 2019) it might be helpful to look at a prediction of total rides based off the average of Q4 (October - December) from past years.

The average for those 15 months is:

```
avgoctdec = sum(octdec_counts) / len(octdec_counts)
```

```
156170.75
```

Once the average is calculated and added to 2019's missing data, I found that the plot ended up looking like this:



Indego Bike Data: Total Passholder Types 2015-2019

Now, I think it is still important to note that this data and these plots are to be taken with a grain of salt because it is merely a prediction for Q4 of 2019 that was found using an average of data from Q4 of prior years. For example, If we are to look solely at the numbers for Q4 (Oct. - Dec.) from prior years we see 2017 had an abnormal spike in rides during those months, which therefore affected the Q4 2019 average.

```
for x, y in enumerate(octdec_counts):
    print(years[x], y)
```

```
2015 121783
2016 176119
2017 183909
2018 142872
```

The average does not take into account why 2017 had such a crazy spike in Q4 and if that factor causes the spike again in 2019 the data could look much different. The opposite is true as well. Q4 may underperform compared to the average.

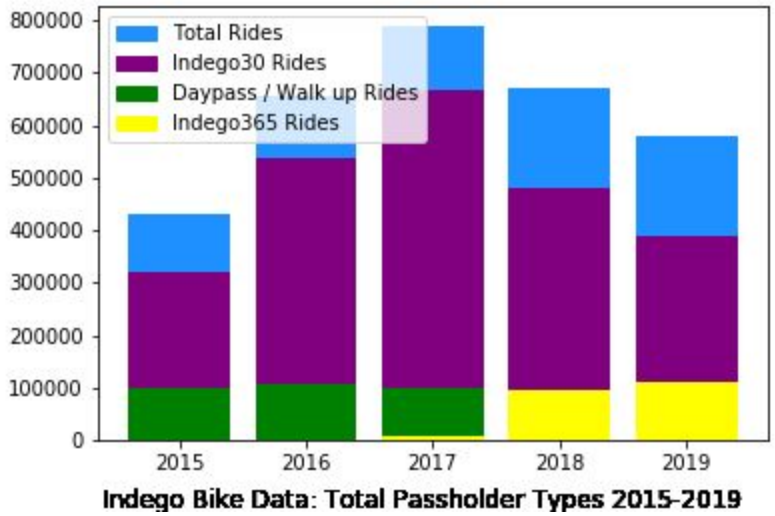
Taking all of this into consideration, I found that I could not accurately predict the Passholder Types for 2019 with the averages, which left me with two options:

1. Remove all Q4 observations from the data
2. Keep Q4's observations and run the analysis with it being known from the get-go that we are missing Q4 2019 data

Since the new year is almost here and Indego will release their Q4 numbers relatively soon, I chose to continue with the second option and will rerun the script once there is a complete data

set for 2019.

The Final Graph has lots of interesting observations on passholder types.



Going back to the original questions I was interested in (what plan is being utilized the most and is the “Indego 365” truly the “best value” based on the Trip Data?) it would seem that “Indego 30 Rides” are year after year, the most popular choice for riders. With the introduction of the “Indego 365” plan in 2017 the plans quickly shifted to an increase in “Indego 365” plans over the next three years.

I am very interested to see how the Q4 2019 data affects “Indego 365” plans compared to “Indego 30” plans. It is my belief that it is very plausible that the “Indego 365” could be the “best value” for riders if the trend increases as predicted year after year.

On a final note, thank you to Indego for making the data public!