# Meghan McGrath
Data Analysis Portfolio

Projects

01	**GameCo**	Video game market analysis to advise strategic planning for future business initiatives
02	**Influenza**	Influenza data analysis to inform staffing needs to upcoming flu seasons
03	**Rockbuster**	Movie rental service analysis to guide new online video strategy
04	**Instacart**	Customer spending habit analysis to inform business strategies
05	**Pig E. Bank**	Customer retention analysis for global banking provider
06	**Airbnb**	Geographical & machine learning analysis for NYC rentals

01
GameCo
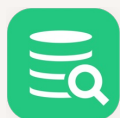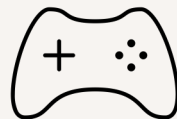
GameCo, a new video game company, wants to use data to inform the development of new games.

Goal: Perform a descriptive analysis of a video game data set to foster a better understanding of how GameCo's new games might fare in the market.

01 Overview

Key Business Questions

1. Are certain types of games more popular than others?
2. What other publishers will likely be the main competitors in certain markets?
3. Have any games decreased or increased in popularity over time?
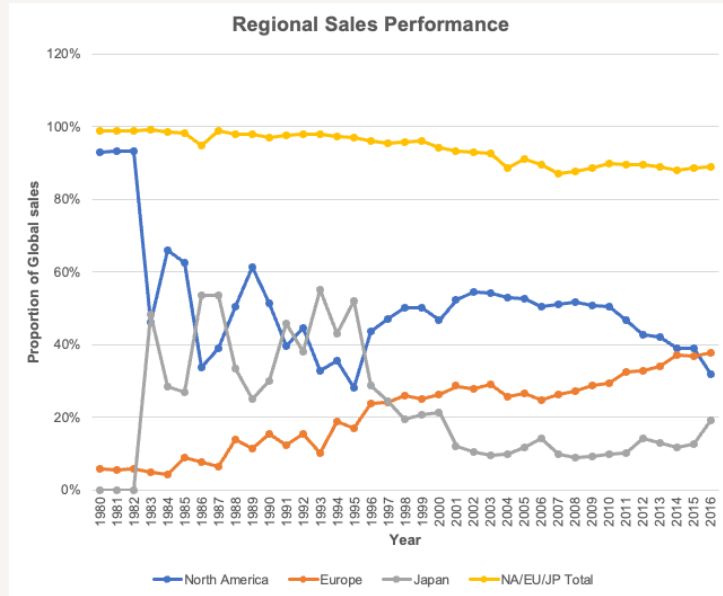4. How have their sales figures varied between geographic regions over time?

Data

Video Game Sales
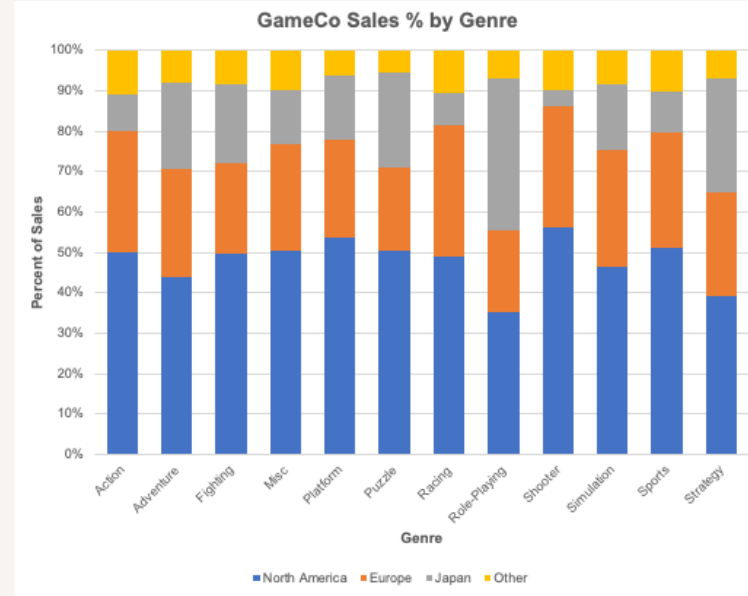Source: [VGChartz Data Set](#)

Tools/Skills

- Excel
- Cleaning data
- Grouping data
- Summarizing data
- Descriptive analysis
- Visualizing data insights in Excel
- Storytelling with data

Findings

Sales in Europe have gradually increased while North American and Japanese sales are volatile and in decline.

Shooter and sports games continue to increase in share while there is a declining interest in platform and puzzle games. There is a consistent low interest in adventure and racing games.

Recommendations

New Insights		
Trouble in North America & success in Europe	**Declining markets still have growing subset categories**	**Opportunity within Emerging Markets**
• North America is a problem region – with the market share decrease in NA, it's left room for other regions to undercut NA and grow their share, with one of the main regions being Europe	• Consumer interests are changing North America (and around the world) • North American sales show that in a declining market, there are still genres that continue to perform well	• New markets are cannibalizing the existing regions' market shares – there is potential to continue their growth with the right strategy.
Recommendations		
Focus on high performing genres in NA	**Shift Resources to Europe**	**Focus on Emerging Markets**
• Allocate more resources and budget to promoting shooter and sports games & pull back on platform and puzzle games due to loss in interest • Alter marketing budgets, reorganize store layouts, and reallocate stock	• The more streamlined resourcing for North America should free up more resources that can be allocated in Europe • With the European population topping NA by 200 million, there is a larger target market aka more potential consumers to capture	• With a new 10% of market share coming from secondary geographical regions, there are entire new markets that can be explored • Source more granular data to identify specifically where these spikes are occurring and what is contributing to them, both internally and externally.

02
Influenza

The United States has an influenza season where more people than usual suffer from the flu. Some people, particularly those in vulnerable populations, develop serious complications and end up in the hospital. Hospitals and clinics need additional staff to adequately treat these extra patients. The medical staffing agency provides this temporary staff.

Goal: Help a medical staffing agency that provides temporary workers to clinics/hospitals to plan for influenza season when additional staff are in high demand. The results will examine trends in influenza and how they can be used to proactively plan for staffing needs across the country.

02 Overview

Key Business Questions

1. What time(s) of year is influenza's impact so great that ancillary staff will be required to tackle additional patient load?
2. What areas of the USA are most likely to demand this extra support?
3. How do vulnerable populations factor into this matter?

Data

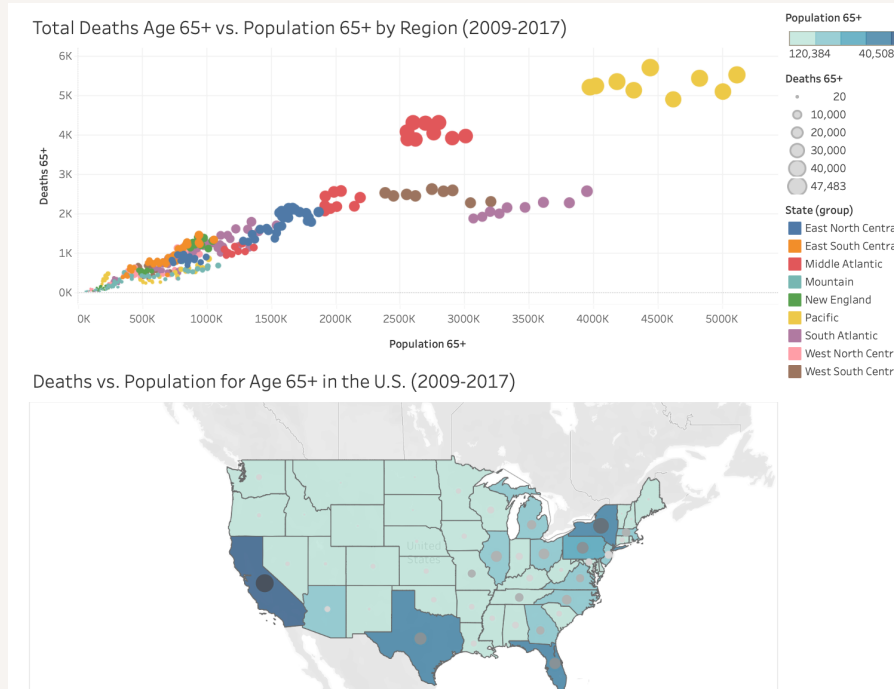**Influenza deaths by geography**
Source: CDC
Data Set

Population data by geography, time, age, and gender
Source: US Census Bureau
Data Set

Tools/Skills

- Excel
- Designing a data research project
- Data profiling and integrity
- Data transformation & integration
- Statistical hypothesis testing
- Data visualization (composition/comparison charts, statistical visualization)
- Spatial & textual analysis
- Forecasting
- Storytelling in Tableau

Findings

Total Deaths Age 65+ vs. Population 65+ by Region (2009-2017)

Deaths vs. Population for Age 65+ in the U.S. (2009-2017)

Relationship between population and deaths among 65+ age group:

- The trendline indicates a positive correlation, meaning that as the 65s+ population increases, so do their deaths.

Findings

Monthly Deaths by Region (2009-2017)
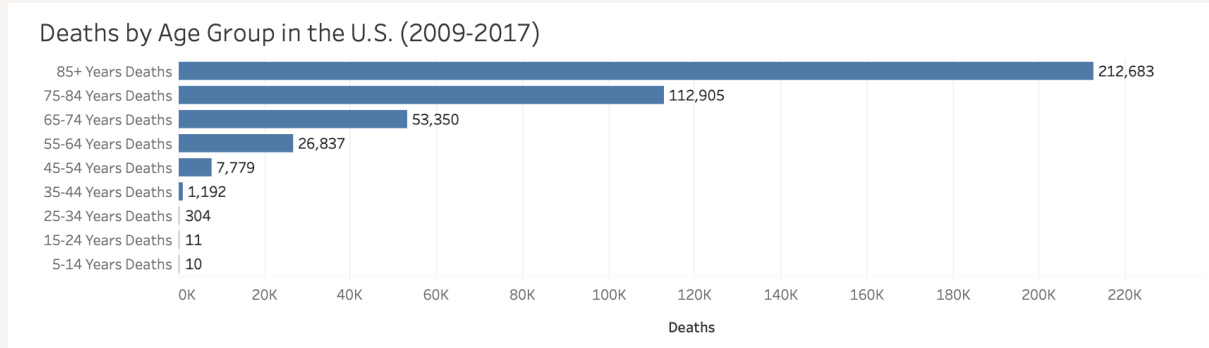

Deaths Forecast by Region (2018-2020)
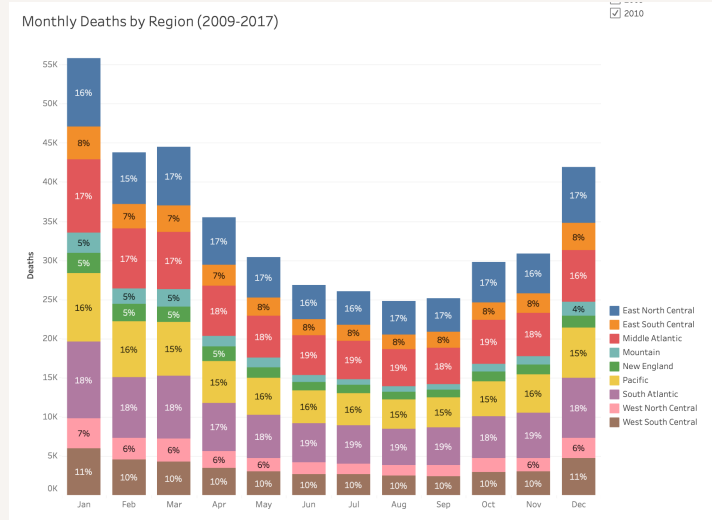
Influenza is seasonal – the highest mortality rates occur during the winter months of November through April.
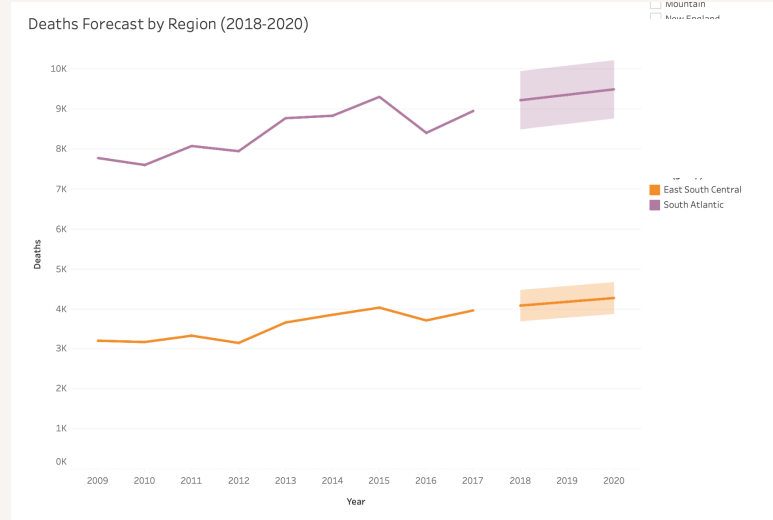
Priority regions for influenza staffing include South Atlantic and East South Central states. The deaths for these regions are forecasted to increase by the greatest rates, by 6% and 7%, respectively.

Recommendations

Deaths by Age Group in the U.S. (2009-2017)

Age Group	Deaths
85+ Years Deaths	212,683
75-84 Years Deaths	112,905
65-74 Years Deaths	53,350
55-64 Years Deaths	26,837
45-54 Years Deaths	7,779
35-44 Years Deaths	1,192
25-34 Years Deaths	304
15-24 Years Deaths	11
5-14 Years Deaths	10

WHERE to send staff: Deaths for populations aged 65 and older make up over 90% of total deaths. The staffing agency will need to prepare to send more medical personally to the following states with the largest vulnerable populations/deaths:
- California, Texas, Florida, North Carolina, Tennessee, Pennsylvania, Illinois, New York, Maine, Michigan

WHEN to send staff: The staffing agency will need to prepare to send more medical personnel during the winter/spring months, November–April as influenza is highly seasonality.

Tableau: found here

03
Rockbuster

Rockbuster Stealth LLC is a movie rental company that used to have stores around the world. Facing stiff competition from streaming services such as Netflix and Amazon Prime, the Rockbuster Stealth management team is planning to use its existing movie licenses to launch an online video rental service to stay competitive.

Goal: Analyze the performance of the current catalogue and develop a launch strategy.

03 Overview

Key Business Questions	Data	Tools/Skills

Key Business Questions

1. Which movies contributed the most/least to revenue gain?
2. What was the average rental duration for all videos?
3. Which countries are Rockbuster customers based in?
4. Where are customers with a high lifetime value based?
5. Do sales figures vary between geographic regions?

Data

Rockbuster Data Set

Tools/Skills

- Relational databases SQL
- Database querying
- Filtering data
- Summarizing & cleaning data
- Joining tables
- Performing subqueries
- Common table expressions
- Presenting SQL results

Findings

Unique Titles	Unique Genres	Average Rental Duration	Average Rental Rate	Average Customer Revenue	Languages
1001	17	4.98 days	$2.98	$102	1
		Min: 3 days Max: 7 days	Min: 0.99 Max: 4.99		

Findings

Top Markets: India, China, U.S & Japan

Top Markets

Country	Revenue
India	$6,035
China	$5,251
United States	$3,685
Japan	$3,123
Mexico	$2,985
Brazil	$2,919
Russian Federation	$2,766
Philippines	$2,220
Turkey	$1,498
Indonesia	$1,353

Customer Count: 14 — 60
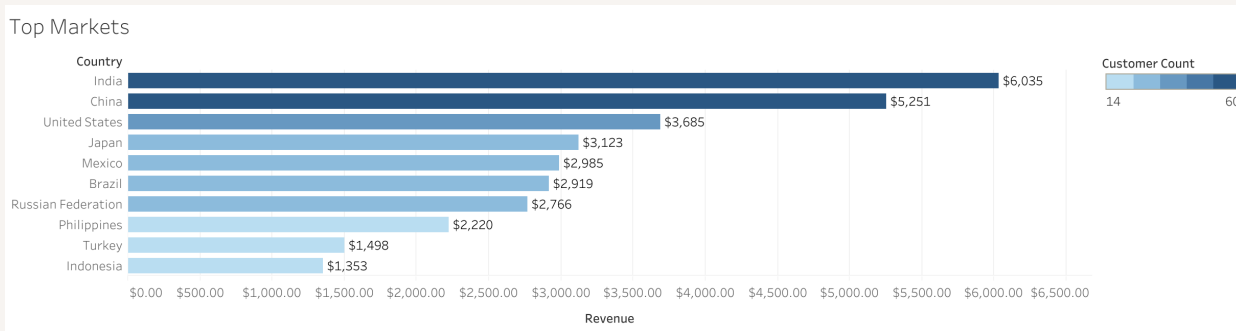
Customer Count (Total vs. Top) by Country

Top Customer Count: 0, 3
All Customer Count: 7 — 60

© 2024 Mapbox © OpenStreetMap

Findings

Top 10 Titles by Revenue

Title	Total Revenue
Telegraph Voyage	$216
Zorro Ark	$200
Wife Turn	$199
Innocent Usual	$192
Hustler Party	$191
Saturday Lambs	$191
Titans Jerk	$187
Harry Idaho	$178
Torque Bound	$170
Dogma Family	$169

Count of Rentals: 22 — 28

Top 10 Genres by Revenue

Genre	Total Revenue
Drama	$2,735
Games	$2,391
Sports	$2,363
Comedy	$2,173
New	$1,977
Foreign	$1,913
Sci-Fi	$1,874
Horror	$1,829
Action	$1,804
Travel	$1,773

Rental Cnt: 278 — 433

Top Titles:
- Telegraph Voyage
- Zorro Ark
- Wife Turn
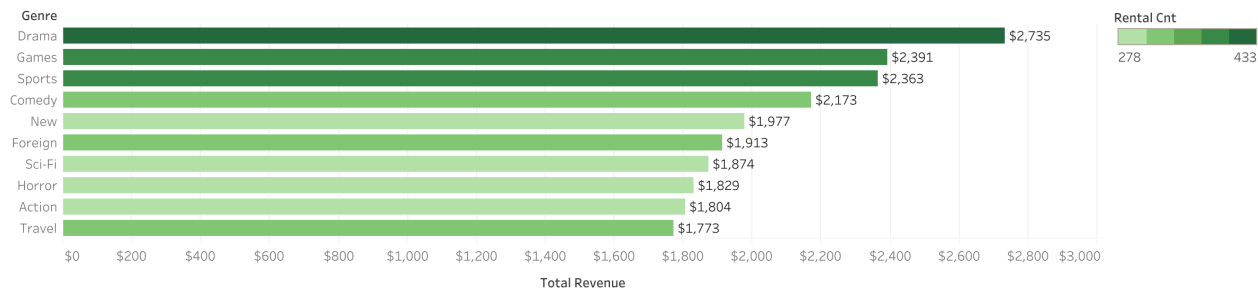
Top Genres:
- Drama
- Games
- Sports

Recommendations

Diversify

- Diversify language offering
- Focus on high performing markets: India, China and the U.S.
- Diversify offerings to include titles in Hindi, Chinese and Japanese

Expand

- Expand library for popular genres
- Increase rental fees for popular genres
- Conduct further analysis to determine title additions to the drama, games and sports categories

Reimagine

- Adjust budget to support locations that are struggling and boost those that are thriving
- Consider introducing penalties for extended rental duration

Data Dictionary: found here
Full Analysis: found here
GitHub: found here

04

Instacart

Instacart, an online grocery store that operates through an app, wants to uncover more information about their sales patterns. The Instacart stakeholders are most interested in the variety of customers in their database along with their purchasing behaviors to implement a targeted marketing strategy.

Goal: Perform a data and exploratory analysis to derive insights and suggest strategies for better segmentation based on the provided criteria. Inform what this strategy might look like to ensure Instacart targets the right customer profiles with the appropriate products.

04 Overview

Key Business Questions

1. When do users spend the most money?
2. What are the most popular products?
3. What's the distribution among users in terms of brand loyalty? Are there differences in ordering habits based on a customer's loyalty status?
4. Are there differences in ordering habits based on a user's region?
5. Is there a connection between age and family status in terms of ordering habits?
6. What different classifications does the demographic information suggest?
7. What differences are there in the ordering habits of different customer profiles?

Data

CareerFoundry Data Sets:
Customers Data Set

Instacart Data Sets:
Data Dictionary
Citation: "The Instacart Online Grocery Shopping Dataset 2017", Accessed from www.instacart.com/datasets/grocery-shopping-2017 via Kaggle on <date>.

Tools/Skills

- Python
- Data wrangling & subsetting
- Data consistency
- Data combining & exporting
- Deriving new variables
- Grouping data
- Aggregating variables
- Data visualization with Python
- Reporting in Excel

Findings
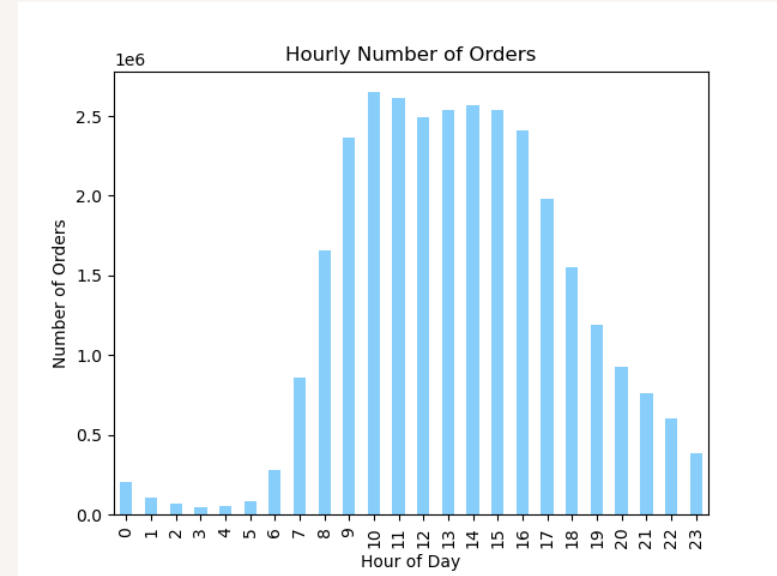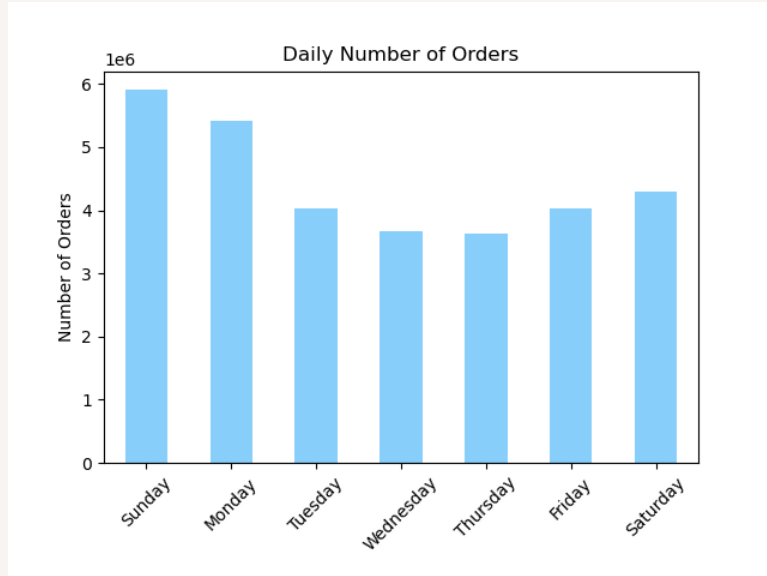
Busiest hours of the day: 9 AM to 5PM
**Busiest days of the week:** Saturday – Monday

Findings – Geographic
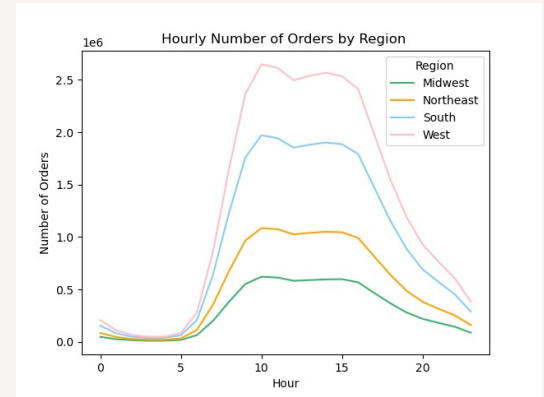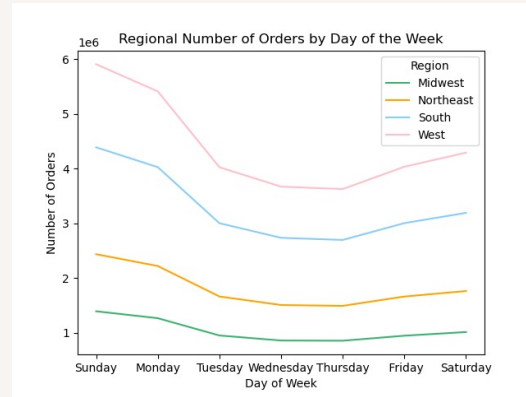
There are **no notable differences in purchasing behavior between the 4 regions** – for all regions, the average price, orders by day and orders by hour remain the same. Instacart does not need to segment its strategy by region.

Findings – Customer Profiles
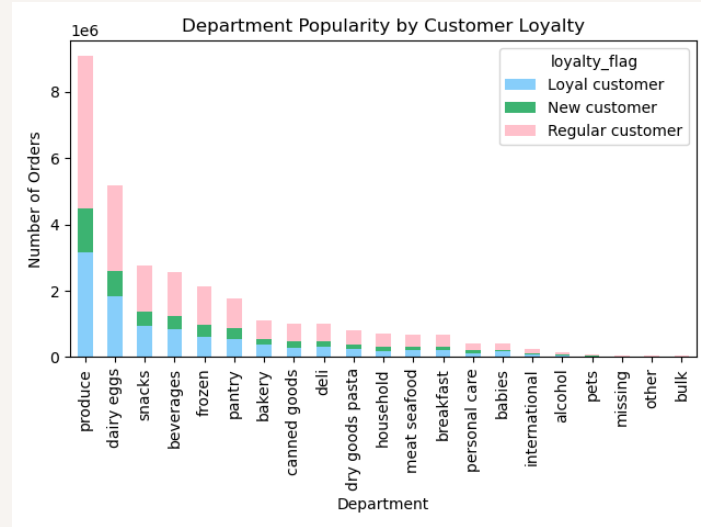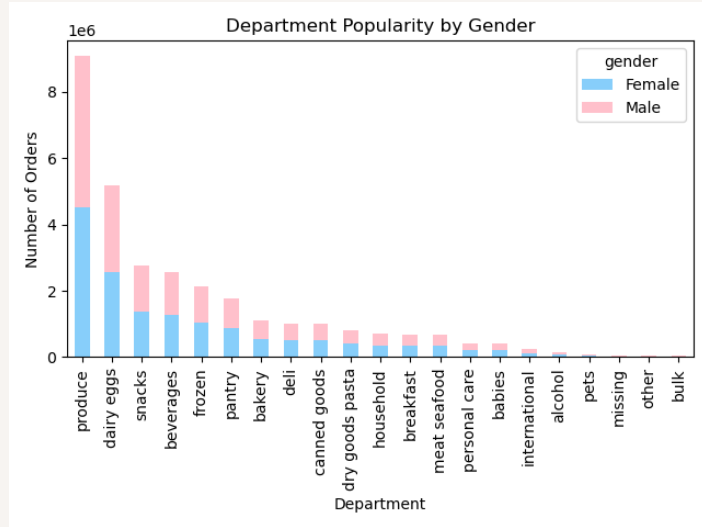
While there are segments that are higher revenue drivers, there is **no significant difference** in what items are being purchased between the different gender and loyalty segments alone.

Findings – Customer Profiles

Customer Profiles and Revenue
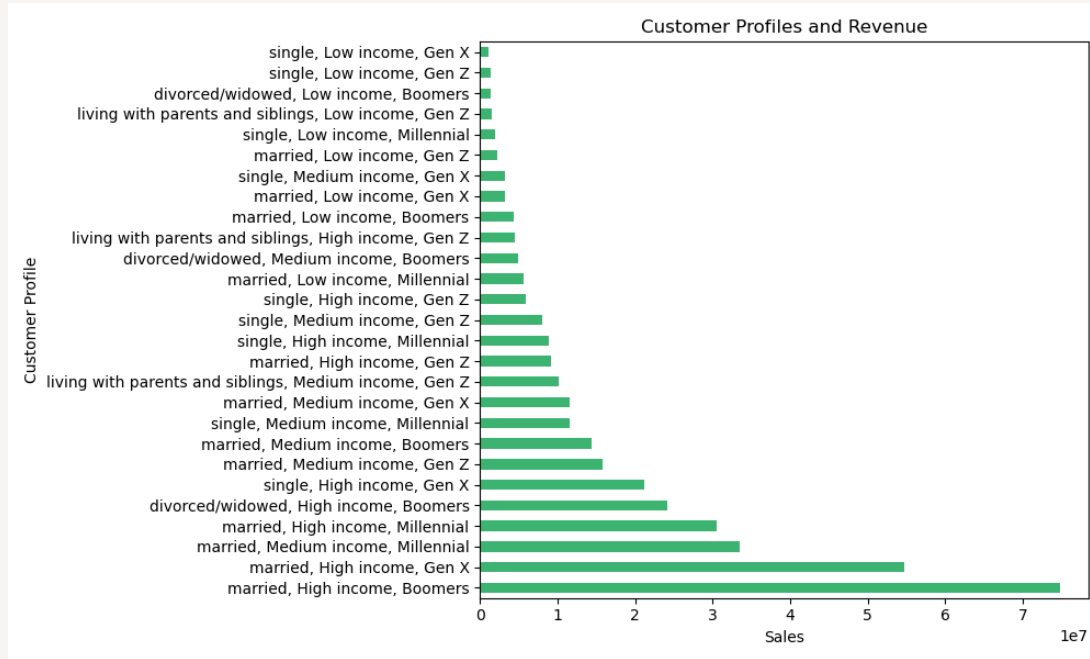
Highest spenders:
- high income, married, Boomers
- high income, married, Gen Xers

Lowest spenders:
- single, low income, Gen X
- Single, low income, Gen Z

Recommendations

- **Schedule ads during the busiest hours of the day**, primarily 9 AM to 5PM, and refrain from advertising during the evening/early morning hours when consumers are asleep. Focus on advertising during the weekend into the new week (Saturday - Monday) as consumers are most active.

- **Focus on upselling regular customers to become loyal customers and maintaining loyal customers through extensive retention marketing** (email, push notifications, re-targeting through PPC, SMS, organic social, etc.).

- **Differentiate marketing strategy among customer profiles:**
  Older, wealthier, family-oriented – advertise via mail/print vs. online, suggest pricier or larger serving add-ons before checkout
  Younger, lower-income singles - push notifications for sales on previously bought or similar items, offer deals on produce with a closer sell-by-date or minor defects, suggest low-priced shelf-stable staples

- **Capitalize on popular departments,** produce and dairy/eggs: these items need to be bought frequently as they have a short shelf life. With these users being on the site more frequently, Instacart should target them more often and continue to collect data to improve their business.

Full Analysis: [found here](#)
GitHub: [found here](#)

05
Pig E. Bank

Pig E. Bank is a well-known global bank looking for anti-money laundering and customer retention support. The Pig E. Bank stakeholders are most interested in assessments of client risk and transaction risk, as well as reporting on metrics.

Goal: Highlight risk and help build/optimize models that assist the bank in running their compliance and retention programs more efficiently.

03 Overview

Key Business Questions	Data	Tools/Skills
1. What are the leading factors leading to client loss? 2. Do certain customers have a higher risk of leaving the bank than others? 3. Are there any data privacy, data security or data bias issues?	**Career Foundry Datasets:** Pig E. Bank's customer data	• Big data • Data ethics – data bias and security & privacy • Data mining • Predictive analysis • Time series analysis & forecasting • GitHub

Findings

This predictive model determines the **probability of customers leaving the bank**
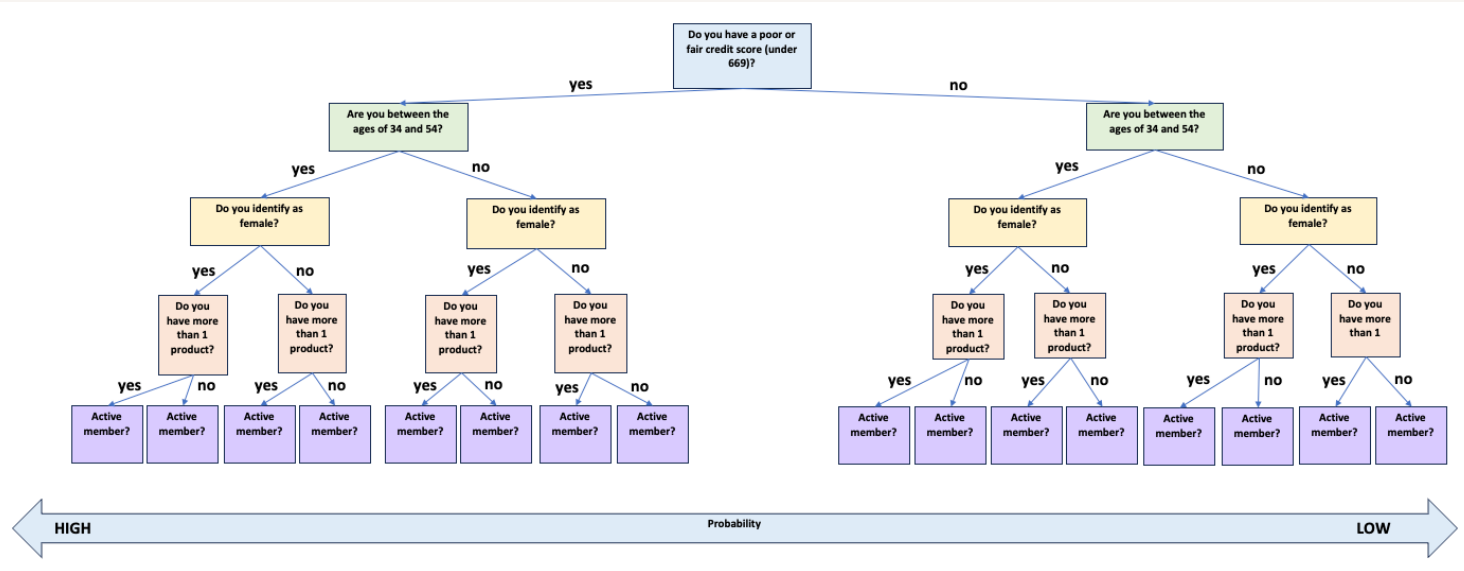
The analysis determined the **highest risk factors** as:
- credit score, age, number of products, member activity status

Recommendations

Promote financial health

Higher credit scores decrease exit rates – provide financial advice and implement customer-centric features (ex. fee-free overdrafts, early paycheck access)

Incentivize activity

Inactive customers are more at risk of leaving – encourage customer activity by expanding number of products or offering promotions (ex. referral bonuses)

Customer Feedback

Better understand customer banking needs through regular surveys and interviews

Biases that may be present in the exercise:
- **Collection** – no explanation of how data has been collected as the team leader for Pig E. Bank has provided stale data from 2017
- **Sample** - the sample provided lacks detail on which ATM transactions are included in the dataset meaning we cannot be sure this is a random sample representative of the total
- **Exclusion** – no explanation of how data provided has been cleaned
- **Measurement** – lack of standardization in identifying transactions as "positive" or "negative"

06
Airbnb

Since 2008, Airbnb has revolutionized travel by offering guests and hosts the opportunity to explore unique and personalized experiences around the world. This dataset provides insights into listing activity and metrics within New York City.

Goal: Make predictions and draw conclusions to further enhance understanding of the Airbnb market in this dynamic urban landscape.

03 Overview

Key Business Questions

1. Is there evidence of spatial autocorrelation in listing prices or booking patterns? Are nearby listings more like each other in terms of pricing and popularity than listings farther apart?

2. Are there any specific features that correlate with higher ratings or booking frequency?

3. Can we forecast future demand or pricing trends based on historical data using time series analysis techniques?

Data

Open-Sourced Airbnb Data: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

Tools/Skills

- Geographical visualizations with Python
- Supervised machine learning – regression
- Unsupervised machine learning – clustering
- Sourcing and analyzing time series data
- Creating data dashboards
- GitHub

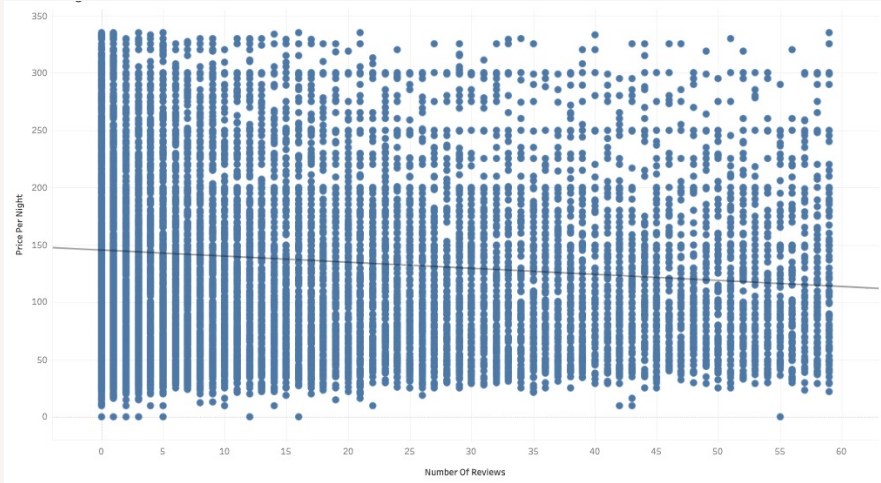# Preliminary Analysis
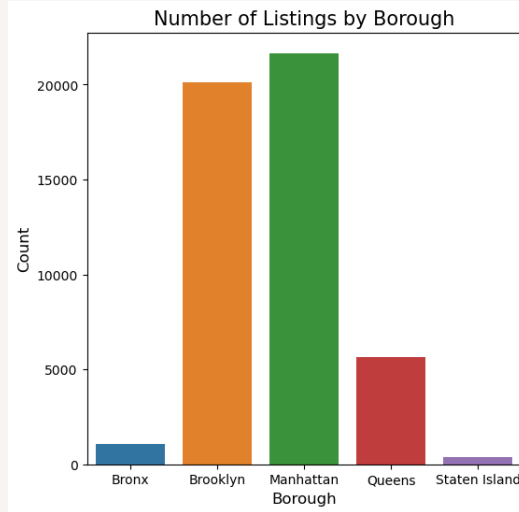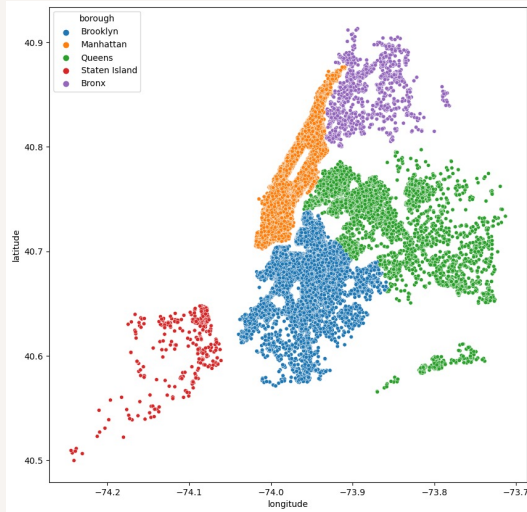
Geospatial Analysis:
As expected, Manhattan (the borough with the most tourist/visitor attractions) has the highest number of Airbnb listings with Brooklyn as a close follower.
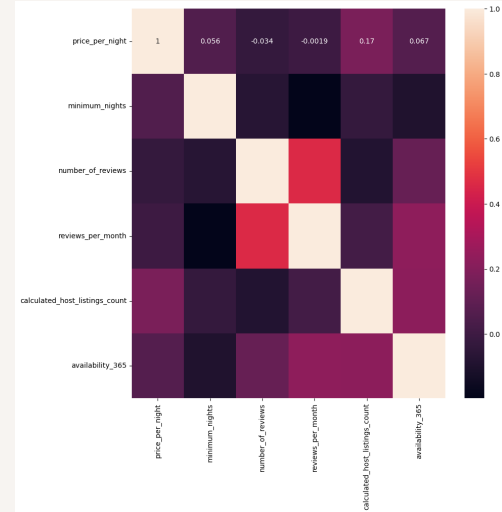
Exploratory Analysis:
Minimal correlation between any of the variables

Findings

Price Per Night vs. Number of Reviews by Borough

Hypothesis: The higher the number of reviews, the higher the price per night. To test this hypothesis, a linear regression was conducted.

With an r-squared value of 0.01, we can deduce that this model is not a good fit.
With a p-value of near 0, we can reject our null hypothesis - there is no relationship between the number of reviews and the price per night of an Airbnb.
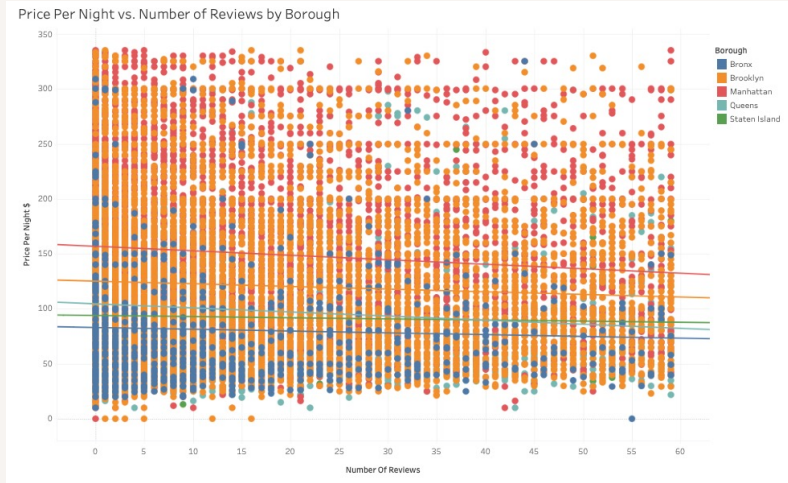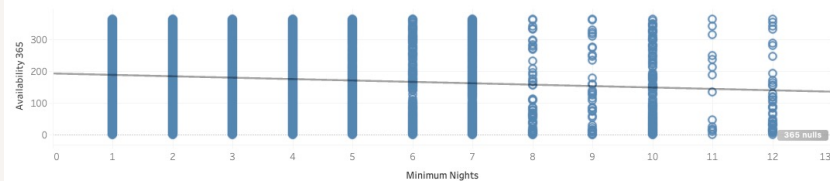
We see the same lack of fit when looking at each borough individually.

However, there are 2 boroughs where we **cannot** reject the null hypothesis due to the p-value: Staten Island (.66) and the Bronx (0.23).

Findings

Regression Analysis - Availability vs. Minimum Nights TOTAL

Regression Analysis - Availability vs. Minimum Nights by Borough

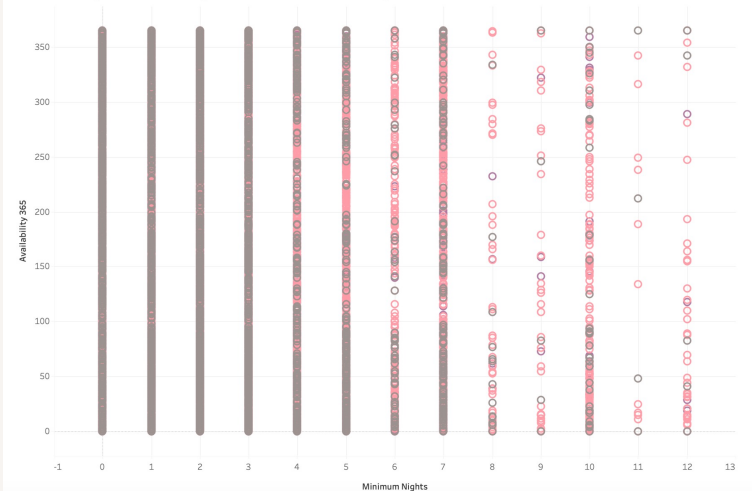Cluster Analysis - Availability vs. Number of Host Listings
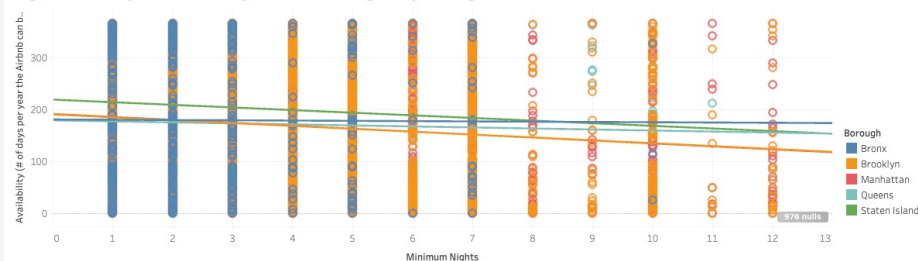
Hypothesis tested via linear regressions: The higher the minimum nights required to book, the higher the year-round availability.

With an r-squared value of 0.01, we can deduce that this model is not a good fit.
With a p-value of near 0, we can reject our null hypothesis – there is no relationship between the number of reviews and the price per night of an Airbnb when looking at all boroughs.
However, looking into individual boroughs, there are 3 boroughs where we **cannot** reject the null hypothesis due to the p-value: Staten Island (0.32), the Bronx (0.87) and Queens (0.22).

The dark purple cluster has the highest availability (i.e. days the Airbnb is available to book online) and the highest number of total Airbnb listings per host – both significantly larger than the pink and purple clusters.

Recommendations

Hypothesis testing: Both were disproved by a linear regression analysis where the r-squared and p-values were both insignificant.
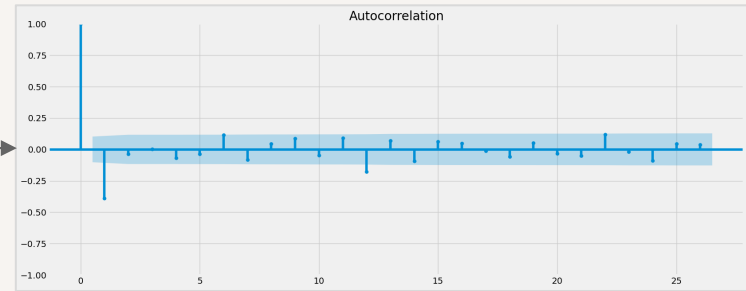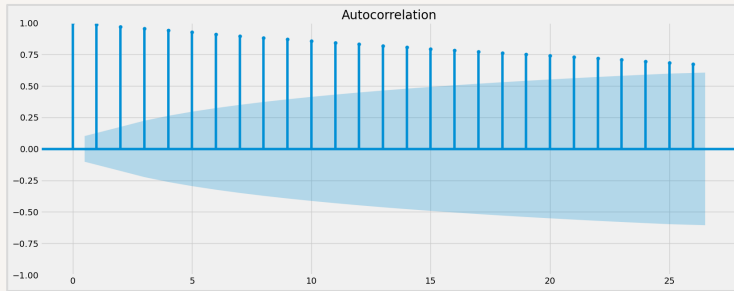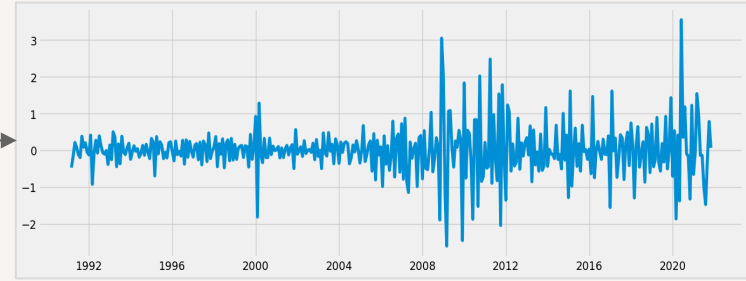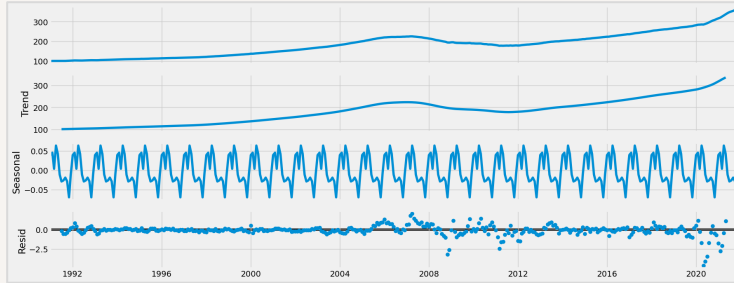
1. The higher the number of reviews, the higher price per night
2. The higher the minimum nights required to book, the higher the year-round availability

Next Steps:
- Analyze the individual boroughs where we could NOT reject the null hypothesis:
  - Hypothesis 1: Staten Island and the Bronx
  - Hypothesis 2: Staten Island, the Bronx and Queens

Bonus – Time Series

Conducted a **Dickey–Fuller test** on real estate data to check for stationarity & performed a round of differencing to check the data's autocorrelations.

Thank you!

Any questions?

meghanmcgrath97@gmail.com

(973) 309 – 5186