Team Notes

**Links:**

Slides: [PAND-O-RAMA](https://docs.google.com/presentation/d/17GM5uTyxJLzCg_lRaqT9b2IzwyCeK1Y-QrRTiDBBgus/edit#slide=id.g2ac28b1027f_1_135)

Example presentation Anthony shared: <https://www.youtube.com/watch?v=qCXJBX72e-Y&ab_channel=AnthonyTaylor>

Main goal of the project

* Act of doing the analysis: meaning each member gets experience coding and participating in all areas of the project

# Session 1 - 12.20.23

* + Group Name
    - Pand-o-rama
  + Choose Git Master - Brandon
    - Create GitHub
    - jezadahlgren
    - mmccanse
    - sonjabaro
  + Short list of topics
    - AirBnB
    - Attrition
  + Action items:
    - Look at data before next class (optional)

## timeline/milestones

* + - Session 2 - 12.21.23
      * Data source exploration
      * Question forming/hypothesis forming
      * finalize topic
      * Cleansing strategy
        + roles/responsibilities
    - Session 3 - 1.3.24 (Added all notes and created hypothesizes, graphs, columns to discuss, PowerPoint, etc. for next class to share)
      * Cleansing / manipulation
      * Data exploration
      * question/hypothesis forming
      * Analysis plan
        + What statistics will we use?
        + What do we want to predict?
    - Session 4 - 1.4.24
      * question/hypothesis forming
      * visualization
    - Session 5 - 1.8.24
      * visualization
      * Presentation - google slides - no more than 10 slides
      * Documentation
    - Session 6 - 1.10
      * presentation

**Group norms:**

* How do we work together?
  + Be able to understand all the parts, talk through work other team members have done
  + Share process, talk through work
  + Show and tell at start of work sessions (time box)
  + Set goals/agenda for working sessions at start (time box)
* Git Hub
  + Individuals write code
  + During group check in - bring up any items that may need to be edited in code
  + People review others code - peer review is really helpful/important
* Data questions
  + Question dump in this doc, people dump questions there, helps to see other’s questions
* Roles/responsibilities?
* Best way to communicate?
  + Slack/email?
* Strengths
  + Sonja
    - Market assessment, what problem are we trying to solve
    - Business Case, requirements docs
    - Presentation/sales skills
  + Meredith
    - Note taking/Documentation
    - Interested in doing part of the coding
  + Jennifer
    - Experience w/ survey research/analysis
    - presentation/sales skills
  + Brandon
    - Deep dive into details/understanding
    - Time management - deadlines
* What’s important?
  + End result - working document that can be shared with others in job search
  + Evidence of market assessment skills relating to/using AI
  + Opportunity to go back to roots with skills / explore/understand analysis (Jen)
  + Good setup to make career change

## Hypotheses Questions

1. Certain aspects of listing influence overall rating. What are they? - Jennifer
   * Choose 3-4 factors, plot each against overall rating, see what most closely correlates
     + Number of listings per host - any correlation to ratings? Do ratings go down the more listings same host has?
     + Price as relates to rating
     + Cleaning fee as relates to rating, % of total rate
     + What is the average rating?
     + Listed features
     + Availability 90-60 days
2. Certain aspects of listing influence price. What are they? - Meredith, Brandon
   1. Neighborhood
   2. Property type / room type
   3. Accommodates (max capacity of listing)
   4. Number of reviews
3. What impacts a quality host? - Sonja
   * Host neighborhood vs. listing neighborhood - if host doesn’t live in neighborhood with listing, is there correlation with rating or usage?
   * Number of cancellations
   * Type of host (superhost or no)
   * Length of time being a host
   * Volume of descriptive text about the property
     + Create new vars length of description: small, medium, large for each: name, summary, space, description
   * Most recent calendar update
   * Number of reviews

## Potential cleansing strategies

* Group by neighborhood and and look at average data by neighborhood
* Suggest delete rows that have blanks in the 7 “Review Scores” columns
* Any records with no review score - remove
  + Remove unnecessary columns - add columns to delete below
    - List of columns here, see ‘keep-discard’ tab: [Denver 23 Data Sample](https://docs.google.com/spreadsheets/d/1EsDRNp8q2Lk0we0nodQHWSRxPUuntn0hEqqBU-NtL20/edit#gid=1657741026)

## Analysis strategy

* We need to remember to verify the license info on the data
* JZD: Suggest the comparison be between City
  + We should merge the data or have one dataframe that has each city listed - this will help with analysis such as correlations or any predictions that we attempt (larger n=size means more power to the analysis and better prediction)
  + MM: I agree, may need to add column with city location in each DF before merge.
  + SB: are we saying merge 2017 and 2023? Do we need to combine? This seems like extra work and isn’t necessary if we’re looking to understand the factors for a rating, and if we want to compare by city, if they vary. (Along with average price etc. If we do merge, would the time gap an issue? Or we could evaluate what we see after doing some basic analysis/Q&A on the years independently? Or do we just use a slice to analyze the data for each set and compare, but have them as a single DF?
* JZD: Question for Group: do we need to have a separate data requirements document for submission?
  + MM: not sure I understand the question
  + SB: I think that would help us summarize for the ppt - I don’t think its a requirement for the submission, but would be an added item. Basically documents what decision points we had, ie (sample size, continuity of data, fields, etc). And importantly what questions we wanted to ask.
* MM: Maybe makes sense to divide the coding tasks like this?:
  + SB - Question - are we changing from each person doing their own city? I thought we were going to all work the same process, but on the different data set? This allows us to be working on the project together, while each getting to get the experience. Also would be a way to do a “code review” for our parts. - Not saying we are changing, was just a thought. As I looked at the data, and based on Jen’s suggestion to combine it, wondered if it makes more sense to divide it by hypothesis.
  + Clean data, remove unnecessary columns, combine 4 data sets into 1 - (person 1, maybe during 1/3/23 session?)
  + Maybe work on the below during 1/4/23 session, and people finish up over the weekend?
    - Rating (person 1)
    - Price (person 2)
    - Host (person 3)
    - Start preparing slides (person 4)
  + Before/During 1/8/23 session
    - Create slides (2 people)
    - Solidify documentation README (1 person)
    - Finish any coding (1 person)
    - Practice presentation

## Visualization strategy

* Heat map
* plot(subplots=True) to see numerous columns plotted separately in same figure
* .corr function to compute correlations
* JZD: Use city on the horizontal axis
* SB: Question - are we comparing city to city? If yes, I would suggest we do the analysis of each city separately first - so I’m assuming the graph being referenced above is for the city v city analysis - ie what factors where most impactful to a score - we could do a scatter with a quadrant plot
* JZD: One page of the slide deck should incorporate all exploratory or nifty findings which will leave more room for the advanced items i.e., .corr, if there are more than that we should include in appendix
  + Slide 1: Title
  + Slide 2: Data (data types and where we got it), Methodology (analysis plan, sample sizes, etc.), Hypothesis
  + Slide 3: Exploratory data or supporting data
  + Slides: Hypothesis 1 - Overall Rating
    - Jennifer
  + Slides: Hypothesis 2 - Price
    - Meredith and Brandon
  + Slides: Hypothesis 3 - Host
    - Sonja
  + Slide 7: Hold for additional corresponding data
  + Slide 8: Hold for additional corresponding data
  + Slide 9: Insights, recommendations, limitations of the study
  + Slide 10: Questions?

## Ideas:

Here are some key aspects and recommendations on how to use this data:

1. **Listing Details**: The dataset includes basic details like listing URL, name, summary, and description. Analyze these for common themes or keywords that could indicate popular features in Boston listings.

2. **Location Data**: You have information on neighborhoods and geolocations. This can be used to analyze spatial distribution of listings, identify popular areas, and examine pricing patterns across different neighborhoods.

3. **Pricing Information**: The dataset likely contains pricing details. Examine how pricing varies with location, type of property, number of bedrooms, or other amenities. This could provide insights into the market dynamics in Boston.

4. **Availability and Booking Patterns**: Data on availability and previous booking patterns (if available) can be used to understand demand trends and peak tourism periods in Boston.

5. **Host Information**: Details about hosts, such as the number of listings they have and their response rate, can be insightful for understanding the market from the supply side.

6. **Reviews and Ratings**: Analyzing reviews and ratings could reveal insights into customer satisfaction and what factors most significantly impact guest experience.

7. **Amenities and Features**: Understanding what amenities are offered and their frequency can help identify what makes a listing more appealing in the Boston market.

8. **Cancellation Policies and Other Rules**: This could be interesting to explore how strict or flexible policies affect booking rates or customer satisfaction.

To conduct a thorough EDA, you should start with data cleaning and preprocessing, such as handling missing values, converting data types if necessary, and normalizing text data. After that, employ descriptive statistics and visualization techniques to uncover trends, patterns, and anomalies in the data. Advanced analyses could include correlation studies, hypothesis testing, or even predictive modeling if you're looking to forecast future trends based on this data. ​​

Top of Form

## Python Tools:

Dependencies

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import scipy as sp

import scipy.stats as sts

import scipy.stats as st

import sklearn.datasets as dta

import requests

import json

import os

from dotenv import load\_dotenv

from census import Census

from pathlib import Path

# Data type conversions and math:

Variables = int, float, bool, str, len

Methods = str., .sum(), .sub(), .mul(), .div(), .mean(), .median(), .mode(),

Functions = max, min, sum, len

Examples: Adding up all listings.

# Formatting

New line = \n

Decmials = .2f(depends on decimal point placement)

Lowercase: {input\_string.lower()}

Uppercase: {input\_string.upper()}

Title case: {input\_string.title()}

Capitalize: {input\_string.capitalize()}

Indexing: [2:4], [1:], [:2]

Add to an end of a list: .append()

Remove: .remove()

Delete: del

Examples: Update values in columns to any specific. Make data more presentable.

Indexing

.set\_index()

.reset.index(), drop=True(if you don't want the index to show)

.reindex()

Print

f-string, str, and .format()

f-string

name = "Alice"

age = 30

print(f"My name is {name} and I am {age} years old.")

str

name = "Bob"

age = 25

output\_str = "My name is " + name + " and I am " + str(age) + " years old."

print(output\_str)

.format()

name = "Charlie"

age = 35

output\_str = "My name is {} and I am {} years old.".format(name, age)

print(output\_str)

Update .csv, .json, html and excel files in pandas

encoding="ISO-8859-1"

sep=';'

delimiter='\t'

index=False

header=True

index\_cols='date'

parse\_dates=['date']

low\_memory=False

normalize=True

Examples: setting index at first including anything to separate (., ;, \_) to import data easier

Update columns

.rename(columns = '')

.sort\_values() with ascending=True(by default) or ascending=False, also by=''

.dropna(), how='any', how='all'

,replace()

.value\_counts()

.fillna()

.astype()

.drop\_duplicates()

.str.replace()

.drop()

.to\_list()

Verify columns

.dtypes

.count()

.unique()

.nunique()

.describe()

.loc

.iloc

Filtering: df.loc[df["Utility"] == "Electricity"]

Subsetting: df[[]]

.isnull()

.isna()

.columns

Show dataframes

display()

print()

df

.head()

.tail()

.info()

Apply Functions

def

.apply()

Data Collecting and Cleaning

.copy()

Updating rows, columns, and dataframes

pd.concat()

df.join()

pd.merge()

df.groupby()

df.groupby().agg()

pd.cut

pd.pivot\_table()

.melt()

.pivot()

Datetime

pd.to\_datetime, include errors='coerce'

APIs

requests.get

.json()

json.dumps()

pd.json\_normalize()

load\_dotenv()

os.getenv

Visualization

plt.plot(), by default, is a line plot

plt.box()

plt.bar()

plt.scatter()

plt.title

plt.xlabel

plt.ylabel

plt.tight\_layout()

plt.legend()

plt.show()

fig1, ax1 = plt.subplots()

ax1.set\_title()

ax1.set\_ylabel()

ax1.plot()

sns.boxplot()

sns.histplot()

sns.barplot()

sns.heatmap()

sns.countplot()

sns.relplot()

sns.lineplot()

sns.jointplot()

Example: Visualize data with all listings or trends within each city.

Example: Use scatterplot to show regression within data (import matplotlib.pyplot as plt)

Example: Use heatmap to show correlation within data (import matplotlib.pyplot as plt and import seaborn as sns). Heatmaps are only available with seaborn

Visualization enchancements

bins=

xlim=

ylim=

figsize=

title=

xticks=

yticks=

color=

rot=

xlabel=

ylabel=

linewidth=

marker=

label=

alpha=

align=

facecolors=

edgecolors=

rotation=

Statistic Methods

.std()

.var()

.corr()

Examples: best maps to use to plot statistical data:

Standard Deviation (std()) and Variance (var()):

Box Plots: Box plots can show the spread of data and highlight outliers.

Histograms: Histograms can provide insights into the distribution of data, with the standard deviation influencing the shape.

Bar Charts: Bar charts can be used to display the standard deviation or variance for different categories or groups within your dataset.

Correlation (corr()):

Correlation Heatmap: A heatmap, as discussed earlier, is a common way to visualize correlations between variables. Seaborn's sns.heatmap() with the corr() matrix is a good choice for this purpose.

Pair Plot: A pair plot can provide scatterplots of variables against each other, along with histograms or KDEs along the diagonal. The correlation between variables is evident in the scatterplots.

## Jen notes:

#Sample code to create a new variable based upon length of comment - checked with chatgpt not with actual data yet

# Dataframe = DF

# Example of defining the buckets for small, medium, and heavy amounts of text

small\_threshold = 50

medium\_threshold = 150

# Create a new column 'text\_amount' based on the length of the 'description'

df['text\_length'] = df['description'].apply(len)

df['text\_amount'] = pd.cut(df['text\_length'], bins=[0, small\_threshold, medium\_threshold, float('inf')],

labels=['small', 'medium', 'heavy'], right=False)

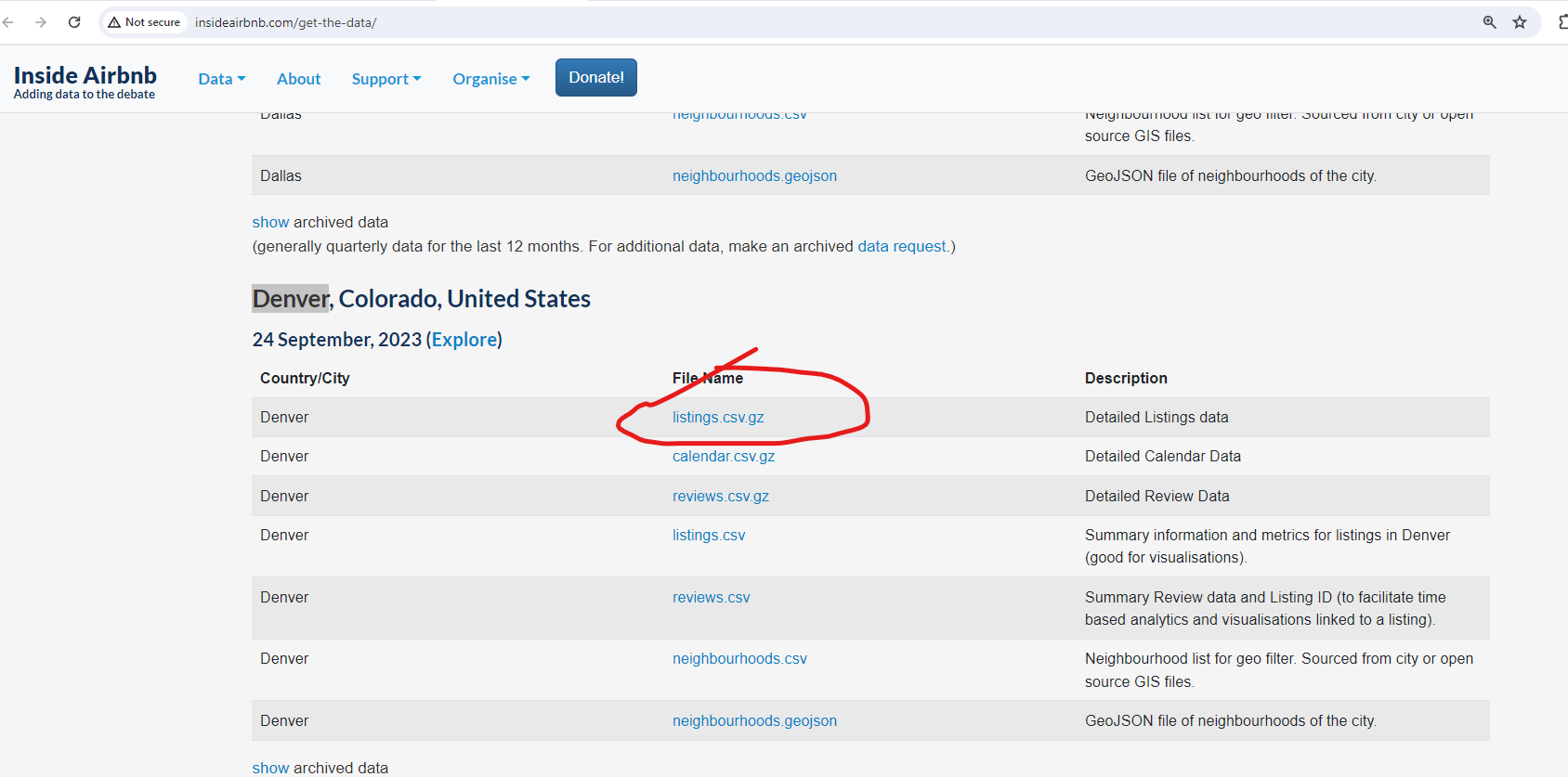
# Drop the 'text\_length' column if we don’t want to keep it

df = df.drop('text\_length', axis=1)

## Meredith notes:

* Was looking more closely at the data, and realized we actually can access the more current 2023 data from this site (<http://insideairbnb.com/get-the-data/>).
  + The .gz format means its been condensed, but I figured out how to read it directly into a Python notebook
  + denver\_all\_listings\_df = pd.read\_csv('DenverBigData.crdownload', compression='gzip')
  + Do we want to use this instead? It has mostly the same data as the 2017 set, but includes additional info, and some of the data presented in more usable formate (ex: superhost is it’s own column with t/f data in 2023 data)
  + If we do, it’s the “detailed listings” data in screenshot below:
  + The data dictionary that Jen shared is actually for this data (columns headers the same, missing some)

| **city** | **2017 number of listings** | **2023 number of listings** |
| --- | --- | --- |
| Portland | 3539 | 4846 |
| Boston | 3382 | 4033 |
| Nashville | 3212 | 8584 |
| Denver | 2479 | 5388 |



Sonja Qs -

1. Nashville has a bunch more rows (about 2x) than the other data sets - I don’t think that’s an issue as we’re looking at them individually first to see what the factors are, but wanted to check.
2. Datatypes - I see int, float & object - wouldn’t the comments be string? (my view is concatenated rn)
3. Let’s discuss the columns - which are needed to answer are questions?
4. What is the basic info we must have to do the analysis? While more info is great - I think our first pass should be somewhat narrow and we can extend if we want to
5. Data cleansing -
   1. Dropping empties: I’m not sure we want to drop rows - need to check the numbers with missing data in the items we want to do analysis - are there other techniques to use?
   2. Which columns do we agree aren’t needed? (discuss as group, thoughts are - listing url, scrape id, last scraped, source, ?picture?, host url,host name, host thumbnail, host picture url,, license, ) (question - is the host neighborhood about where they live? What relevance,?)
   3. Host has profile pic - keep? A Y or N value may influence??
   4. What is difference between neighbour cleansed, neighbourhood group cleansed?
   5. What is diff with max-max-nights and max nights, same for min - repeats several times - do we need the maximum nights info?
   6. availability columns - do we need this? How do we understand
   7. Check out the reviews columns - do we understand these? Do we assume or state we are assuming the review scores rating is the comprehensive rating and the others contribute? How can we investigate this further?there is also the review scores value as well.
   8. What is the meaning of review\_scores\_accuracy?
   9. Are there renames or converting of columns needed at all?
   10. Should we look at the rating distribution (the overall rating) and possibly group/bin the data based on that? - might help with the analysis

# Session 2 - 12.21.23

* Set up GitHub branches
* Retrieve data
* Look at data
* Brainstorm/decide hypothesis questions?
* Final decision on topic
* Action items for break
  + Cleansing strategy
    - Each person explore data, have opinions about how/what to cleanse
    - Add notes [here](#_4ej46d8dum42)
  + Analyzing strategy
    - Add notes [here](#_xt7fr8cdh3p5)

# Session 3 - 1.3.24

## Team Reset - expectations

During class the team came together to reset our expectations as well as what we needed to do to move forward on this project.

1. We have 3 hypotheses that can be found here: [hypotheses](#_iz7i7j26kq7z)
   1. Using the combined data set, explore what factors influence:
      1. Price
      2. Host
      3. Rating
   2. All group members remain present during group discussions, not work on other things
   3. Each group member will have different parts of the project assigned, and should work on their part.
   4. It’s likely that everyone’s work will be adjusted or changed as we combine it and discuss
   5. It’s expected that we will look at each other’s code in order to learn and also help review each other’s work
   6. Presentation
      1. Each person continue to work on core hypothesis question that they focused on
      2. We will divide up the presentation so each person presents a different part of the project

* GOAL for end of night: Have data cleansed by end of this session
* Action items for next session -
  + try for 2 visualizations, thoughts about what found,etc
  + In Excel - add notes about data cleanup
    - [Denver 23 Data Sample](https://docs.google.com/spreadsheets/d/1EsDRNp8q2Lk0we0nodQHWSRxPUuntn0hEqqBU-NtL20/edit#gid=1657741026&fvid=1194744119)
  + Are we missing anything?
  + Brandon
    - Push updates to github - done
    - Price
  + Sonja
    - Host
  + Jennifer
    - Rating
  + Meredith
    - Price

# Session 4 - 1.4.24

Action items for next session:

List of things for us to provide on Monday (will have time to work during class too):

* Brandon (Hypothesis 2)
  + Histogram
  + Read me updates
* Meredith (H2)
  + Work on intro on slide deck
  + Price graph(sorry can’t remember what you said!)
* Sonja (H3)
  + Slide 3 most important points
* Jennifer (H1)
  + Slide 4 most important points
  + Clean up the notes document

# Session 5 - 1.8.24

* Finalize PPT
* Organize github
* Practice presentation
  + Zoom background? - pandas generated from ChatGPT
  + Wear anything specific? - business casual
* Action items
  + Everyone needs to clean up their comments in code

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Appendix:

Information in this section references the process that the group went through in order to come to a finalized product.

## Topic Brainstorm

* + publicapi/publicapi
  + Use API to build that skill
* **AirBnB - housing - rental rates - etc** 
  + <https://public.opendatasoft.com/explore/dataset/airbnb-listings/information/>
  + Cities that we are concentrating on:
    - Denver
    - Boston
    - Nashville
    - Portland
  + Cleansing strategies

HOLD to delete after we come to consensus that we don’t need more data

* + [**https://www.kaggle.com/datasets/marcosgarcia75/air-bnb-dataset**](https://www.kaggle.com/datasets/marcosgarcia75/air-bnb-dataset)
  + **http://insideairbnb.com/**
  + Zillow API for real estate (zillow data not recommended by Kevin, not super granular)
  + Census data
  + Unhoused data
  + <https://www.kaggle.com/datasets/airbnb/seattle/data>
* **Employee Attrition** 
  + <https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset>
  + <https://www.kaggle.com/datasets/henryshan/2023-data-scientists-salary>
  + <https://www.kaggle.com/datasets/jpmiller/employee-attrition-for-healthcare>
  + <https://www.aihr.com/blog/hr-data-sets-people-analytics/>
  + <https://www.aihr.com/blog/hr-data-sets-people-analytics/>
  + <https://www.shrm.org/resourcesandtools/tools-and-samples/hr-forms/pages/turnover-data-by-industry-region-size.aspx> \*\*\* This one looked promising
  + Unemployment data
* Example projects/code for attrition data
  + <https://www.kaggle.com/code/hidayatzeb/hr-attrition-analysis-for-beginners>
  + https://www.kaggle.com/code/paramarthasengupta/hr-analytics-prediction-why-do-people-resign
* Cleansing strategies
* Added topic - are “blue cities” more dangerous than “red”? (based on using the crime datasets that Kevin referenced) We could explore the validity of politician’s claims that democratic leaning/majority cities/states are more crime ridden than others….
* Added topic - do cities with more migrants have more crime? (playing off claims/headlines that immigrants increase crime)
* Added topic - using <https://developer.spotify.com/documentation/web-api>; we could explore listening trends by region and see if they correlate to other factors, or have changed through time.

Suggest looking for the data sources we know are out there and fitting a topic to them…

## Hypothesis Dump

* While unemployment metrics indicate a strong economy in terms of low numbers of unemployed (currently at 3.7% as of November 2023 US BLS), is it healthy when we look at attrition and employer churn? Does the unemployment rate mask other issues about the current state of working in America? Do we see any regional trends? How do we compare with other countries?
  + Potential data sources/sets - US BLS has unemployment, hourly wage, productivity data that could be compared to professional society and other data sets (IBM etc)
* For Airbnb - do we see use of airbnbs increasing or decreasing and is there a correlation to economic factors? Anecdotal reports is that the once hot Airbnb market both as a host and a renter has cooled #

## 

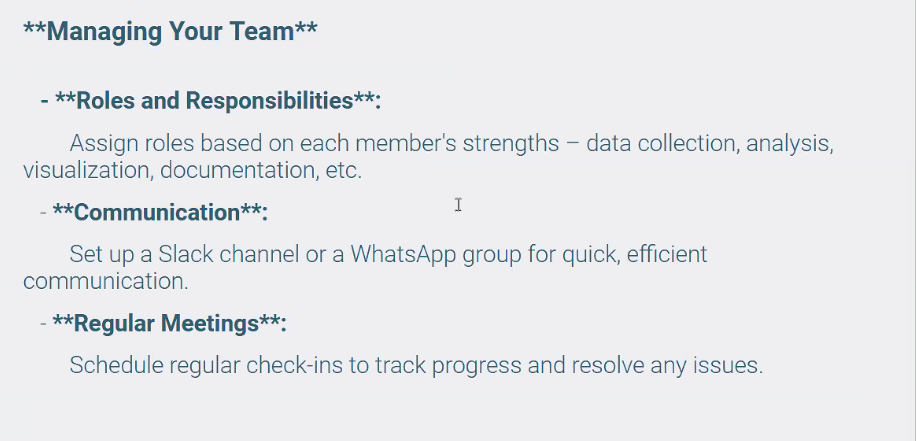
## First Draft of the data dictionary

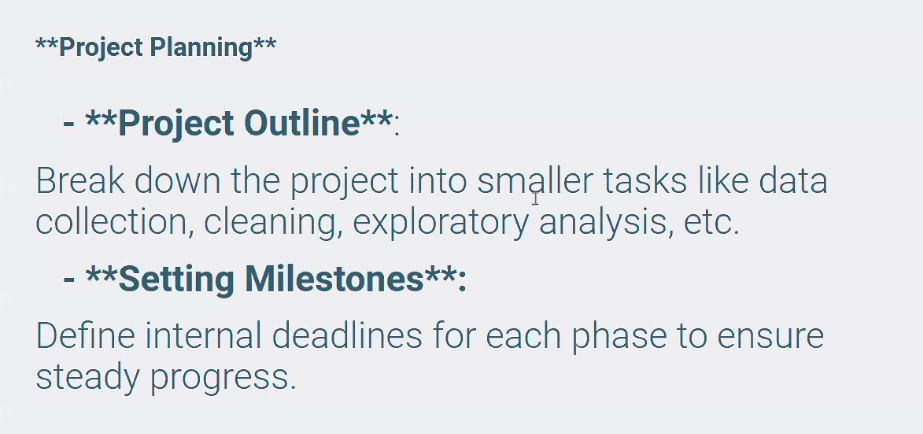
Data Dictionary: https://docs.google.com/spreadsheets/d/1iWCNJcSutYqpULSQHlNyGInUvHg2BoUGoNRIGa6Szc4/edit#gid=1322284596

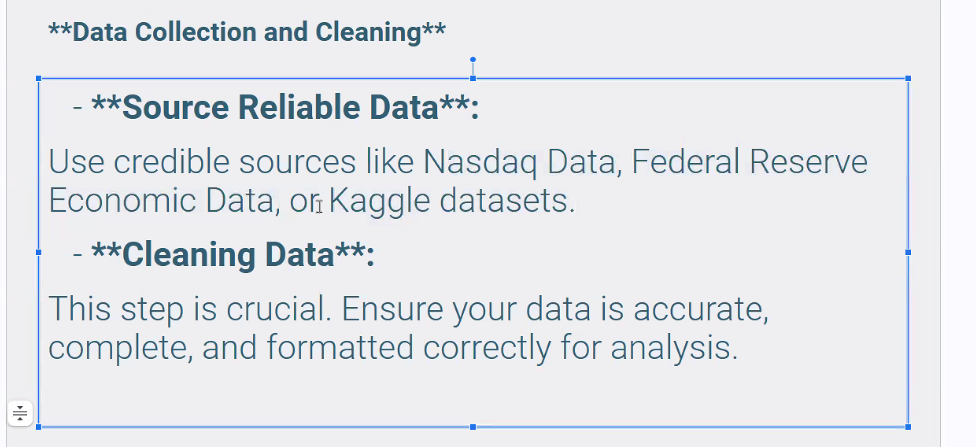
| **Field** | **Type** | **Calculated** | **Description** | **Notes about variables:** | **Variable type: classification, independent, dependent** | **Analysis** |
| --- | --- | --- | --- | --- | --- | --- |
| id | integer |  | Airbnb's unique identifier for the listing | Keep | classification |  |
| listing\_url | text | y |  | Delete |  |  |
| scrape\_id | bigint | y | Inside Airbnb "Scrape" this was part of | Delete |  |  |
| last\_scraped | datetime | y | UTC. The date and time this listing was "scraped". | Keep, make certain in correct formate | classification |  |
| source | text |  | One of "neighbourhood search" or "previous scrape". "neighbourhood search" means that the listing was found by searching the city, while "previous scrape" means that the listing was seen in another scrape performed in the last 65 days, and the listing was confirmed to be still available on the Airbnb site. | Delete |  |  |
| name | text |  | Name of the listing | Delete for Portland  Delete for Denver | Up to host to create a name data to messy |  |
| description | text |  | Detailed description of the listing | Keep, create new variable based upon length of description  Portland has: summary, space, description, neighborhood overview, notes, transit, | Independent | host |
| neighborhood\_overview | text |  | Host's description of the neighbourhood | Keep, create new variable based upon length of description - code sample below | Independent | host |
| picture\_url | text |  | URL to the Airbnb hosted regular sized image for the listing | Delete |  |  |
| host\_id | integer |  | Airbnb's unique identifier for the host/user | Keep | Classification maybe Indenpendent |  |
| host\_url | text | y | The Airbnb page for the host | Keep, create new variable based upon length of description? | Independent |  |
| host\_name | text |  | Name of the host. Usually just the first name(s). |  |  |  |
| host\_since | date |  | The date the host/user was created. For hosts that are Airbnb guests this could be the date they registered as a guest. |  | Independent | host |
| host\_location | text |  | The host's self reported location |  | Independent | host, rating |
| host\_about | text |  | Description about the host |  | Independent |  |
| host\_response\_time | text |  | Response options:  A few days or more, within a day, within a few hours, within am hour | Keep | Independent | host |
| host\_response\_rate | indeger |  | 0 to 100 | Keep, check for outliers | Independent | host |
| host\_acceptance\_rate |  |  | That rate at which a host accepts booking requests. | Delete | Check with other cities but no data in Portland | host |
| host\_is\_superhost | boolean [t=true; f=false] |  |  | Variable not in Portland data | Text located in “Features” but must be created | host |
| host\_thumbnail\_url | text |  |  | Delete |  |  |
| host\_picture\_url | text |  |  | Delete |  |  |
| host\_neighbourhood | text |  |  | Delete |  |  |
| host\_listings\_count | text |  | The number of listings the host has (per Airbnb calculations) | Keep | Might be small sample size, could see which host have multiple properties - create a new variable | host |
| host\_total\_listings\_count | text |  | The number of listings the host has (per Airbnb calculations) |  |  | host |
| host\_verifications |  |  |  |  |  |  |
| host\_has\_profile\_pic | boolean [t=true; f=false] |  |  | Delete |  | host |
| host\_identity\_verified | boolean [t=true; f=false] |  |  | Delete |  | host |
| neighbourhood | text |  |  | Delete |  | price |
| neighbourhood\_cleansed | text | y | The neighbourhood as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles. | Delete  Keep (MM) |  | price |
| neighbourhood\_group\_cleansed | text | y | The neighbourhood group as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles. | Delete |  |  |
| latitude | numeric |  | Uses the World Geodetic System (WGS84) projection for latitude and longitude. | Delete |  |  |
| longitude | numeric |  | Uses the World Geodetic System (WGS84) projection for latitude and longitude. | Delete |  |  |
| property\_type | text |  | Self selected property type. Hotels and Bed and Breakfasts are described as such by their hosts in this field | Response Options: Other  Apartment  House  Bed & Breakfast  Loft  Guesthouse  Hostel  Townhouse  Condominium  Yurt  Bungalow  Cabin  Boat  Tent  Boutique hotel  Hut  Camper/RV  Villa  Guest suite  Earth House  Dorm  Cave  Treehouse | Independent. We will need to group these to use them Suggest running a freq and grouping based upon top three to four housing types and grouping rest into other. |  |
| room\_type | text |  | [Entire home/apt|Private room|Shared room|Hotel]  All homes are grouped into the following three room types:  Entire place  Private room  Shared room  Entire place  Entire places are best if you're seeking a home away from home. With an entire place, you'll have the whole space to yourself. This usually includes a bedroom, a bathroom, a kitchen, and a separate, dedicated entrance. Hosts should note in the description if they'll be on the property or not (ex: "Host occupies first floor of the home"), and provide further details on the listing.  Private rooms  Private rooms are great for when you prefer a little privacy, and still value a local connection. When you book a private room, you'll have your own private room for sleeping and may share some spaces with others. You might need to walk through indoor spaces that another host or guest may occupy to get to your room.  Shared rooms  Shared rooms are for when you don't mind sharing a space with others. When you book a shared room, you'll be sleeping in a space that is shared with others and share the entire space with other people. Shared rooms are popular among flexible travelers looking for new friends and budget-friendly stays. | <https://www.airbnb.com/help/article/5/what-does-the-room-type-of-a-listing-mean>  Response options: Shared room, Private room, Entire home/apt | Indepenent |  |
| accommodates | integer |  | The maximum capacity of the listing | Keep | Independent. Look for outliers i.e., 0 or more than 16? |  |
| bathrooms | numeric |  | The number of bathrooms in the listing | Keep | Indenpenent. Look for outliers |  |
| bathrooms\_text | string |  | The number of bathrooms in the listing.  On the Airbnb web-site, the bathrooms field has evolved from a number to a textual description. For older scrapes, bathrooms is used. |  |  |  |
| bedrooms | integer |  | The number of bedrooms | Keep | Indenpenent. Look for outliers |  |
| beds | integer |  | The number of bed(s) | Keep | Indenpenent. Look for outliers |  |
| amenities | json |  |  |  |  |  |
| price | currency |  | daily price in local currency | Keep | Independent or Dependent |  |
| minimum\_nights | integer |  | minimum number of night stay for the listing (calendar rules may be different) | Keep |  |  |
| maximum\_nights | integer |  | maximum number of night stay for the listing (calendar rules may be different) |  |  |  |
| minimum\_minimum\_nights | integer | y | the smallest minimum\_night value from the calender (looking 365 nights in the future) |  |  |  |
| maximum\_minimum\_nights | integer | y | the largest minimum\_night value from the calender (looking 365 nights in the future) |  |  |  |
| minimum\_maximum\_nights | integer | y | the smallest maximum\_night value from the calender (looking 365 nights in the future) |  |  |  |
| maximum\_maximum\_nights | integer | y | the largest maximum\_night value from the calender (looking 365 nights in the future) |  |  |  |
| minimum\_nights\_avg\_ntm | numeric | y | the average minimum\_night value from the calender (looking 365 nights in the future) |  |  |  |
| maximum\_nights\_avg\_ntm | numeric | y | the average maximum\_night value from the calender (looking 365 nights in the future) |  |  |  |
| calendar\_updated | date |  |  |  |  |  |
| has\_availability | boolean |  | [t=true; f=false] |  |  |  |
| availability\_30 | integer | y | avaliability\_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host. | Keep | Dependent |  |
| availability\_60 | integer | y | avaliability\_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host. | Keep | Dependent |  |
| availability\_90 | integer | y | avaliability\_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host. | Keep | Dependent |  |
| availability\_365 | integer | y | avaliability\_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host. | Keep | Dependent |  |
| calendar\_last\_scraped | date |  |  | Investigate what this is? |  |  |
| number\_of\_reviews | integer |  | The number of reviews the listing has | Check with other cities to see what they have Portland has “reviews per month” | Independent. |  |
| number\_of\_reviews\_ltm | integer | y | The number of reviews the listing has (in the last 12 months) |  |  |  |
| number\_of\_reviews\_l30d | integer | y | The number of reviews the listing has (in the last 30 days) |  |  |  |
| first\_review | date | y | The date of the first/oldest review |  |  |  |
| last\_review | date | y | The date of the last/newest review |  |  |  |
| review\_scores\_rating |  |  |  | Keep | Dependent |  |
| review\_scores\_accuracy |  |  |  | Keep | Independent |  |
| review\_scores\_cleanliness |  |  |  | Keep | Indenpendent |  |
| review\_scores\_checkin |  |  |  | Keep | Independent |  |
| review\_scores\_communication |  |  |  | Keep | Independent |  |
| review\_scores\_location |  |  |  | Keep | Independent |  |
| review\_scores\_value |  |  |  | Keep | Dependent ? |  |
| license | text |  | The licence/permit/registration number |  |  |  |
| instant\_bookable | boolean |  | [t=true; f=false]. Whether the guest can automatically book the listing without the host requiring to accept their booking request. An indicator of a commercial listing. |  |  |  |
| calculated\_host\_listings\_count | integer | y | The number of listings the host has in the current scrape, in the city/region geography. |  |  |  |
| calculated\_host\_listings\_count\_entire\_homes | integer | y | The number of Entire home/apt listings the host has in the current scrape, in the city/region geography |  |  |  |
| calculated\_host\_listings\_count\_private\_rooms | integer | y | The number of Private room listings the host has in the current scrape, in the city/region geography |  |  |  |
| calculated\_host\_listings\_count\_shared\_rooms | integer | y | The number of Shared room listings the host has in the current scrape, in the city/region geography |  |  |  |
| reviews\_per\_month | numeric | y | The number of reviews the listing has over the lifetime of the listing |  |  |  |

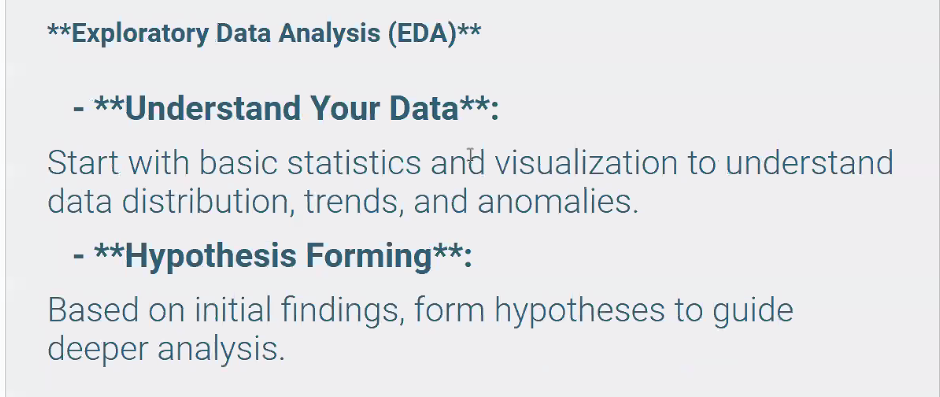
# Slides

+

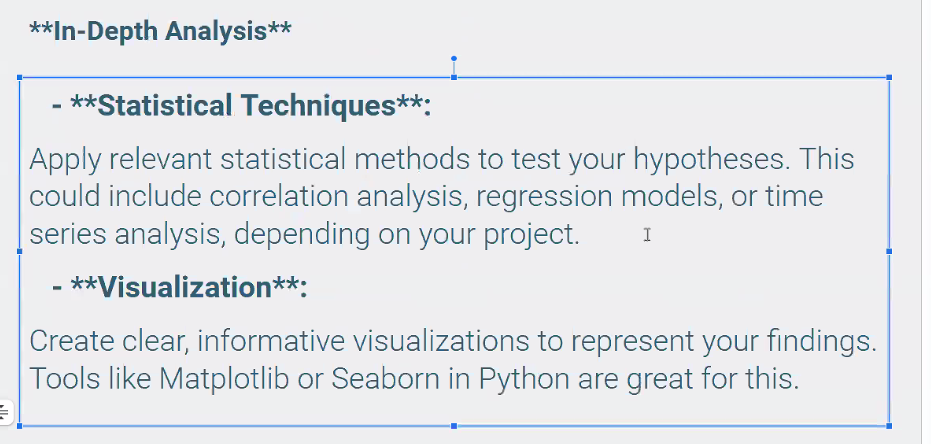




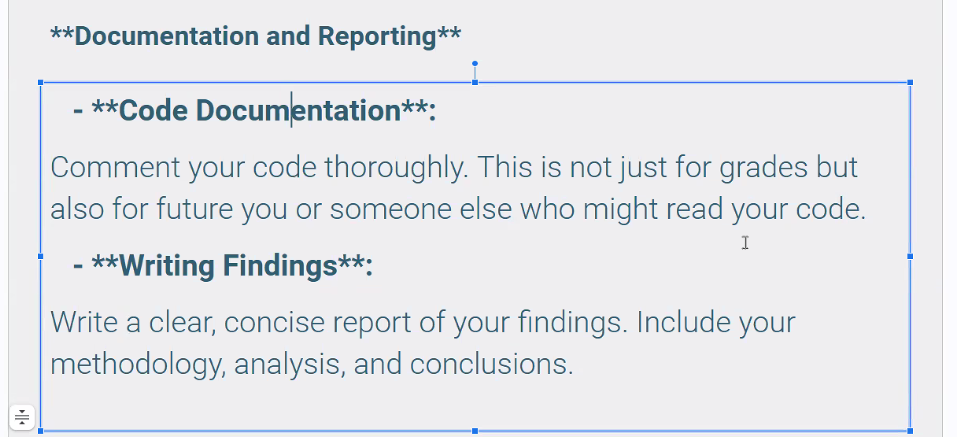


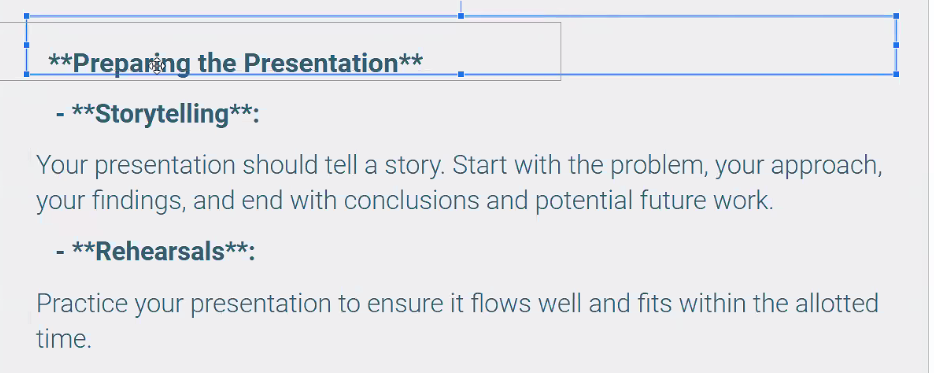


* Hypothesis forming: Look at the data and come up with 3 or 4 questions, use visual data to answer the questions

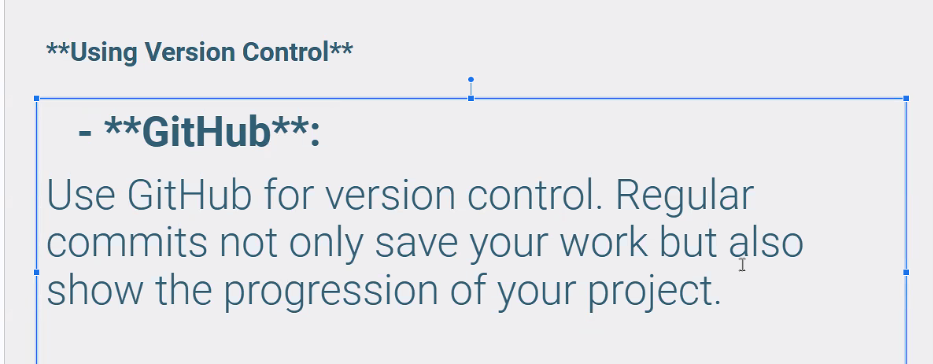


* Try to use Prophet somewhere
* 2-3 visualizations per question





* Presentation is 30 points
* Don’t care about our code, what did we do
* Presentation time: 10 minutes



* Create branch for each person, eventually push into main
* Main branch is only part that gets graded
* Don’t name your files the same thing - everyone include initials in file name

