**My Ideas/Questions for the EDA Project**

1. **Price data from Quandl**

**Business value** – create and test trading strategies to find unique and uncorrelated ones in order to build a portfolio of positive expectancy and relatively low risk equity growth.

**Questions**

Can you come up with a simple day trading strategy that shows positive results across multiple assets which goes home flat every day?

Can you find a strategy that works on uncorrelated markets?

Find what features/metrics of combination of them that predicts future price movement the best.

1. **Football play data or score data with lines**

**Business value** – use data to make better predictions about final scores, ranking of teams, player statistics. Profit from deviations in public lines/fantasy sports betting sites.

**Questions**

What raw data about a team/offense/defense/special teams is best to predict the amount of points scored or surrendered in a given game.

Do key injuries (QB specifically) create overreactions from the public, therefore the lines, and therefore create opportunities?

**Prompt #1 – NYC Subway Data**

**Background** – We will explore the NYC MTA turnstile data set. These data files are from the New York Subway. It tracks the hourly entries and exits to turnstiles by day in the subway system.

**Data** – If you use the raw turnstile data from the [NYC MTA](http://web.mta.info/developers/turnstile.html) it is in a TXT format that closely resembles .csv, but contains extra white space appended to each line. pandas.read\_csv should properly convert the final column “EXITS” to int64, but the column name will contain extra spaces. You are not limited or required to use the raw turnstile data. New York City has a large amount of additional data at: <https://opendata.cityofnewyork.us/>.

**Business Value** – Transportation data has a long history in urban planning and civil engineering, often providing insight into everything from where traffic is most congested to where bike lanes would be most effective at preventing vehicle-bicycle collisions. The potentials for adding business value through transportation data are vast.

An interesting recent case study has been Uber in 2017. It released a website called Movement that uses *Uber’s* data to help urban planners make informed decisions about cities, and it has pledged to make it easier for riders to keep tabs on transit options by displaying real-time public transportation data directly within the Uber app.

The 3 questions below are example prompts for using transportation data to aide business and non-profit decisions.

**Prompt** – Using publicly available data on the NYC MTA in conjunction with any other publicly available datasets, answer one or more of the 3 questions below.

**Question 1** – Can you identify the best areas to canvass for signatures?

**Details** – You are working with a non-profit organization who is trying to raise awareness about women in technology. They want you to identify the best areas to canvass during the day. They will place street teams at the entrances to various subway stations to collect email addresses. You need to help them optimize the placement of street teams. You will want to assume they will canvas on a specific date or month in the future. After your preliminary analysis on the turnstile data set, can you incorporate demographic analysis as well? See Gabriel’s Jupyter notebook for an example that incorporates information about New York City schools. What other demographic information could you find about subway riders that would make them more prone to be interested in the cause of women in technology?

**Question 2** – Can you track how subway usage changes with weather patterns?

**Details** – You are working for a clothing retailer who wants to plan certain pop-up events and flash sales around busy and pleasant weather days. What days are historically busy and pleasant? Help the clothing retailer pick a few days over the next few months that will maximize their exposure and minimize the chance of poor weather.

Using turnstile data is one way to indicate the busyness of a day and looking at monthly weather history is one way to predict the chance of poor weather. In analyzing the busyness of a day, you may want to normalize your data to remove large annual events like the Christmas uptick or the Tennis US Open in August.

**Question 3** – With the plethora of census data, can you identify patterns among various income and ethnic groups?

**Details** – You are working with an advertising firm that wants to target certain demographics at different subway stations. Median income, rent prices, and population density in each neighborhood can be culled from the U.S. Census. What subway stops serve different income and ethnic groups? Does the income and ethnicity of the surrounding area influence the rate of subway use?

**Prompt #2 – NYC Rental Listing Data**

**Background** – [This](https://www.kaggle.com/c/two-sigma-connect-rental-listing-inquiries) dataset comes from Kaggle and its goal is to predict the interest level in New York rental properties based on information from rental listings. The training dataset contains various aspects of rental listing information and your goal is to relate the mostly numerical variables to rental’s interest level. Interest level is simply the number of inquiries a new listing receives.

**Data** –  The data is in JSON format which stands for JavaScript Object Notation and is a lightweight data-interchange format mostly used in web development. JSON is briefly covered in the API section in “Getting and Cleaning Data”.  Don’t worry too much about JSON, but if you work as a web developer you will likely encounter JSON again. It is simply another way to markup data in vein of HTML or XML, except that it is simpler because it is a collection of name/value pairs similar to a dictionary in python. Pandas comes with a convenient pandas.read\_json method that will parse the training data into a pandas DataFrame similar to how you could read a csv or other tabular data format. However, you will notice that the row index is not sequential, and you may want to run reset\_index() to fix this.

Each listing gives the following information:

1. Date of listing creation
2. Number of bathrooms (float values that include half bathrooms)
3. Number of bedrooms
4. Building\_id
5. Description of rental property
6. Display\_address
7. List of features about this apartment
8. Geo location (latitude, longitude)
9. A list of photos of the rental property
10. Price: in USD
11. Street address
12. Interest\_level that this listing generated. Ordinal values of: 'high', 'medium' and 'low'

**Business Value** – In this prompt you have a specific client: RentHop. For more information on their mission see the Kaggle prompt. Having a deeper understanding of what determines a rental’s interest level will help RentHop better handle fraud control, identify potential listing quality issues, and allow owners and agents to better understand renters’ needs and preferences.

**Prompt** – Using Kaggle data in conjunction with any other publicly available datasets, answer one of the three questions below.

**Question 1** –What determines the interest level of a property?

**Details** – Do properties with lower rental prices show higher interest levels or vice versa? If there is a price vs. interest level trend, is this trend prevalent when accounting for other factors? Meaning do we see a relationship between interest level and price once we factor in location and number of bedrooms or bathrooms. Can you assign each property to a neighborhood (Soho, West Village, East Village, etc.) using its geolocation and examine the interest level based on neighborhoods?

**Question 2** – How does proximity to Subways influence a rental’s price or interest level?

**Details** – Subway is one of the preferred modes of transportation in NYC. It seems intuitive to expect a correlation between rental price and distance to the closest subway. Is there correlation between price and distance to the nearest subway? What about once you account for the other factors that influence price (such as number of bedrooms and bathrooms)? Do we see a similar relationship between the distance to the nearest subway and interest level?

**Question 3** – What are the most common features of high, medium, and low interest properties? Does the description of the property influence interest level?

**Details** – This question requires you to explore the messier columns of the raw dataset. For example, the feature column is an array of string keywords that describe features of the listing and the description column is a free string that verbosely describes the listing. Note that the description column contains html line breaks (<br />) that you may need to remove.

The Counter class in python might be useful for tallying the occurrence of different features or words for each interest level. A counter object in python supports convenient and rapid tallies. It contains a most\_common method that you could use to look at the most common features associated with low, medium, and high interest level apartments. Similarly, you could tally the words that appear in the descriptions of low, medium, and high interest level properties and compare their rate of use.

**Prompt #3 – Chicago Crime Data**

**Background** – This dataset comes from [this](https://www.kaggle.com/currie32/crimes-in-chicago) Kaggle competition. “This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. “

**Data** – The data is contained in CSV files. Because there are years of data here, the scope of your analysis is largely up to you. We recommend that you choose to do an analysis of a one to three year time period, but you are welcome to do an expansive analysis. Pandas read\_csv is ill formed to read some of the lines in the data. If it does not eliminate a large swath of the data you may use the error\_bad\_lines option on read\_csv to skip lines.

**Public Value** – Analyzing this data has value to law enforcement in terms of resource planning and criminologists looking for explanatory factors and trends in crime. The scope of analysis here could be vast and range from simple trends or hotspot visualizations to sophisticated geospatial predictive analytics that correlate the geographic features of a city with past criminal events to predict future crime. While you are unlikely to build a system on the scope of the precogs in Minority Report, this kind of data has an opportunity to assist those actively working to combat, curtail, and understand crime.

**Prompt** – Using Kaggle data in conjunction with any other publicly available datasets, answer one or more of the three questions below. The City of Chicago has a large trove of public data including [geographic data](https://data.cityofchicago.org/browse?tags=gis) mapping everything from the location of police and fire stations to bike racks, and [socioeconomic data](https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2) that includes the unemployment rates of different neighborhoods.

**Question 1** – How does crime change over the year? Does it have a seasonality component? Is the trend the same across each neighborhood?

**Details** – Chicago’s crime rate has been front page news of recent. August 2016 marked the most violent month Chicago had recorded in over two decades and Spike Lee brought the issue of Chicago crime to Hollywood in his film Chi-Raq. With this dataset can you visualize the rate of Chicago crime over time? Is there a seasonality component to crime? Does crime happen more in certain months/years compared to others? The **community area column** indicates area where the incident occurred and Chicago has 77 [community areas](https://data.cityofchicago.org/d/cauq-8yn6). Have certain areas seen larger fluctuations or increases in criminal events? It may be useful to control for the number of residents of each community and look at the rate of crime per resident. Understanding the seasonality of crime or identifying where crime rates are increasing will help law enforcement plan efforts to combat crime.

**Question 2** – What community areas have the highest number of criminal events? What are the demographics of high and low crime community areas? **Details** – This question definitely overlaps some with Question 1, but elicits involving demographic information. Unemployment, low incomes, and crowded housing are often indicated as predictors of crime. Do the high crime areas of Chicago demonstrate this commonly accepted paradigm?

**Question 3** – Does the presence of certain geographic features curtail or attract criminal events? Can past locations of criminal events predict future events?

**Details** – This question teases at more complicated criminal prediction models. Does the presence of a police station within a community area decrease the rate of criminal activity? Do liquor stores increase the rate of criminal activity? Do the locations of criminal events of the past week correlate with the locations of criminal events this week?  Each of these questions attempts to get beyond the larger trends discussed in question 1 and 2 and instead look at crime at a more micro level. This kind of analysis would help law enforcement plan weekly patrols given last week’s events and trends.

**Prompt #4 – IMDB Data**

**Background** – In this project, you will explore a dataset that contains information about movies, including ratings, budget, gross revenue and other attributes. This dataset comes from [this](https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset) Kaggle competition. The dataset author scraped [imdb.com](http://imdb.com/) and wondered: “How can we tell the greatness of a movie before it is released in cinema?”. It is also worth noting that there is a public API for pulling information about movies: <http://www.omdbapi.com/>. If you plan to work with this API see the API section in “Getting and Cleaning Data” and google for tutorials for working specifically with the Open Movie Database. Additionally, predicting the gross revenue of a movie is a project we include in our regression problem set. Check Gabriel’s Jupyter notebook there for more information if you plan on scrapping your own data from movie websites.

**Data** – The data should be easily parsed using pandas.read\_csv. I’d recommend looking at the important notes section on Kaggle as well. It includes a note about some of the movies using Korean currency which could complicate your analysis.

**Business Value** – Your goal is to understand what makes movies financially successful, critically acclaimed, or profitable. This task is extremely important as it can help a studio decide which titles to fund for production, how much to bid on produced movies, when to release a title, how much to invest in marketing and PR, etc. This information is most useful before a title is released, but it is still very valuable after the movie is already released to the public (for example it can affect additional marketing spent or how much a studio should negotiate with on-demand streaming companies for “second window” streaming rights).

**Prompt** – Using Kaggle data in conjunction with any other publicly available datasets, answer one or more of the three questions below.

**Question 1** – How does gross revenue vary for genre, content rating (R, PG13, etc.), Facebook likes, budget, cast total likes, IMDB likes, etc.?  In general, how are these variables related to each other? For example, do social media likes align with critic’s views?

**Details** – These kinds of questions influence how much a movie studio might want to spend on social media marketing, on actors, on their budget, or in reaching out to critics to review movies. With evidence from data, decision makers are better able to understand how to meet their objectives. The column genre represents a list of genres associated with the movie in a string format separated by | . Write code to parse each text string into a binary vector with 1s representing the presence of a genre and 0s the absence, and add it to the DataFrame as additional columns. This may help you explore gross revenue as a function of genre. Many of the other features are numerical and lend themselves to scatterplots.

**Question 2** – What famous actors and directors have the highest budget to gross ratio?

**Details** – When a studio spends money on a big name they expect it to greatly increase their revenue, but is this always the case? After selecting actors and directors that appear frequently in the dataset, who is the most worth it? Do these actors and directors consistently appear in one genre or are they profitable across various genres?

**Question 3** – What movies where the biggest flops? Are there factors that are associated with flops?

**Details** – Given both budget and gross data can you quantify this? Remember the note that some movies may be in different currencies.

**Prompt #5 – Stack Overflow Developer Survey**

**Background** – Every year, Stack Overflow conducts a massive survey of people on the site, covering all sorts of information like programming languages, salary, code style and various other information. This year, they amassed more than 64,000 responses fielded from 213 countries.

**Data** – Find the dataset [on Kaggle](https://www.kaggle.com/stackoverflow/so-survey-2017) or [on Stack Overflow](https://insights.stackoverflow.com/survey)

**Business Value** – Think of yourself as the client on this project. Should you go into data science? How can you base your decision on data? Will it increase your salary? Your job satisfaction? Okay so you decided to go into data science… now what tools should you learn? R? Python? Both? What other tools are Data Scientists using? Okay so you learned data science and landed a job. How does your salary compare to others in your field? What about once you account for experience and education?

**Prompt** – Using the stack overflow data in conjunction with any other publicly available datasets, answer one or more of the three questions below. I would encourage you to look at Stack Overflow’s [own analysis of the data](https://insights.stackoverflow.com/survey/2017). There is even a section on bootcamp success! Also check out the stack overflow [trends tool](https://insights.stackoverflow.com/trends?utm_source=so-owned&utm_medium=blog&utm_campaign=trends&utm_content=blog-link&utm_term=state-of-mobile&tags=r%2Cstatistics).

**Question 1** – How do data science salaries compare to other technical career salaries? How is salary effected by years of experience or education?

**Details** – The data contains a USD salary column. Additionally, there is information on geographic location, education level, developer type, etc. Do data scientists make more than other technical professions? Does a data scientist’s salary have different explanatory variables than other careers?

**Question 2** – What are the most popular tools and languages among data scientists?

**Details** – You’re working with Python in K2 because it is one of the most popular data science languages. Does the data backup this assumption? You’re also working with Jupyter notebooks. Are your peers doing the same? Check out the data columns: IDE and HaveWorkedLanguage.

**Question 3** – Where are data scientists working and who are they working for? How career satisfied are Data Scientists?

**Details** – Check out company size, company type, home remote, and various other columns that describe the working conditions of data scientists. Who are you likely to get a job with after finishing K2? How awesome will your job be?

**Prompt #6 – Choose Your Own Topic and Question**