DATA SCIENCE ASSIGNMENT

The goal of this assignment is to demonstrate to us your ability to manage a real-world data set, to communicate and contextualize summary statistics, and, most importantly, your creativity and analytical rigor in generating and testing hypotheses. The assignment is intentionally nonspecific and exploratory in nature.

The data set at hand captures permits issued by the New York City Department Of Buildings (DOB), and can be accessed here. We strongly encourage you to merge the NYC permit data with other data sets to contextualize your investigation. Mindful of the time we are giving you to complete this assignment, we recommend somewhat narrowly focusing your investigation. We're looking for a document that summarizes your investigation and findings.

(Note – all my work can be found in this git repo - https://github.com/aashishdugar/DOB_Permit_Issuance)

DATA SETS

The data sets that I used were the DOB Permit Issuance, the NYC Subway Stations and the NYC Zip Codes data set.

ETL/DATA CLEANING

The first step I took was building raw data-sets. This was done by either downloading them or running a web scrapper using Python's Scrappy and BeautifulSoup package. The 2 data sets I added were the NYC subway stations and NYC most expensive areas and their zip codes.

Then I carried out some cleaning using Jupyter and the Pandas Library on the given DOB Permit Issuance data set. I cleared out some unnecessary columns to create a more focused dataset as per the query.

QUERIES

So, some of the statistical inferences/hypotheses that I had were the below questions –

1) Borough with most building permits cases
2) Borough with the greatest demolitions/rebuild rate, i.e, NB, DM, A1 job type.
3) Best/upcoming residential areas to live in.
4) Hypothesis - Average Time of job completion from date of permit issuance to expiration can give us an idea on the kind of projects that are happening in the area.

PROCESS/INVESTIGATION

Here is a little detailed description of the process/investigation to the above queries and the steps I took to get them. (Results listed below this section) -

1) I used ***Spark*** for this. Since it's a very simple problem, I ran it in a ***PySpark*** shell and just added the values for each sort of occurrence.
   Here we can see that Manhattan has the most permits cases. Given it's got the highest population density this was expected.

2) Here I try to get a statistic for renovation rate of a borough. Of course, this cannot be a true value as each, as every building isn't represented here, and many permits can be linked to a single building. However, we do have a substantial data set which is quite granular (up to street and avenue level) that helps highlights every area of the borough, which is representative of each area improving itself. Out of all the buildings that have filed a permit, this statistic shows the ones that issue new building/demolition permits and A1 permits. We don't account for alterations A2 and A3 as these account for renovations and not a brand new structure as such (Link for information on the codes in the dataset).

   To carry out this statistic, I decided to use ***Pandas*** within a ***Jupyter Notebook*** as I wanted quick visual updates to what the data set would look like on each transformation or operation. The formula I carried out was a simple percentage formula (We can get the total number of permits from the values at statistic (1) in this report.) –

   $$x = \frac{\text{Values where job type is a }'\text{NB}','\text{ DM}'\text{or }'\text{A1}'\text{ per borough}}{\text{Total number of permits issued per borough}} * 100$$

   Here we see that the majority of permits issues for Staten Island are major updates. Manhattan focused primarily on smaller updates and renovations. Being the densest (population to area ratio) borough, it may seem hard to carry out major construction, as compared to its counter parts.
   Taking that out of the equation, it's also evident that Bronx is the oldest and least wealthy borough, so their rate of update is the lowest in the remaining 4.

3) This is a little bit ambitious as an analytic, but it could give a great outcome. Here we try to guess the best upcoming residential neighborhoods in NYC based on the permits being issues for renovations/new buildings etc.

   All the ads for apartments we see online always specify one thing – the distance from the nearest subway station/bus station, or always have one question from everyone - "How close is the nearest subway/bus station?" And this criterion significantly increases the price of an apartment. This, paired with an already existing knowledge of the most expensive neighborhoods in NYC (i.e. because of the rate of the apartments, the class of people living there etc.) can give us an idea of where the best residential places would be, at a granularity of the street/avenue level.

It’s sort of a binary classification, that is if some criteria in the data set is satisfied, then it’s a yes or else no. It involves a crude comparison which can definitely be refined.

I compared the existing data set to one which has a list of subway stations and one which has high-ranked areas with respect to price and quality of living. These areas were compared with the zip codes provided in the permit data set. As for the subway data, we can compare it with the street names provided in the permit data set. After that, just filter out the view to the get the areas that have subway stations and good neighborhoods around them, and we get the final list.

For this, I used ***MapReduce*** to make some comparisons and add some data (can be considered a type of data cleaning) and then jumped back to ***Jupyter*** with ***Pandas***.

4) Here I continued with the ***Pandas*** and ***Jupyter***. I used ***Matplotlib*** to give a visual for the same. This result was left a little incomplete as there is some additional data for it, but its sufficient to give us a lead to a possible analytic.

   The result essentially shows us the average time it takes from the start of the job to the expiration date of the permit (i.e last day of the job, if a job exceeds this time, a re-issued permit is also available in the table). Of course, this just highlights the time period of each permit, the job can be end well before the end of the permit. But we can still rely on this because the length of the permit is directly proportional to the difficulty of the job, i.e., New Building permits are longer than some simple plumbing or electric permit

   Again, this doesn’t highlight that the average time for doing a job is x days y hours, but the length of the permit is x days y hours.

   The statistic was obtained from a subtraction of the expiration date from the job start date to get the list of time periods of each job. We then take the mean of this per borough.

$$x = \frac{\sum(expiration\ date -\ job\ start\ date)}{\text{Permits per borough}}$$

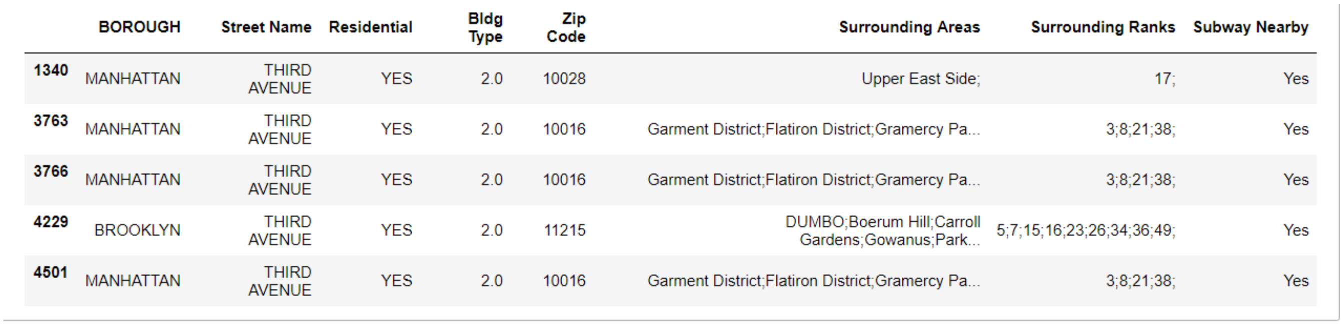
   This can actually lead to various other possible outcomes. It could mean that majority of the jobs in Manhattan are jobs that take lesser time than other boroughs, or maybe the average speed of construction is a lot quicker in Manhattan, or the permit types are just minor changes and nothing major. Whereas in the case of Brooklyn, it could be that it’s the highest because there are a few really major constructions (as seen in the spikes in the histogram in the notebook.) that significantly increase the average days for construction to a much greater value.

RESULTS

1) *[('BROOKLYN', 813934), ('QUEENS', 731242), ('BRONX', 296495), ('STATEN ISLAND', 204365), ('MANHATTAN', 1462213)]*

2) *Manhattan - 8.10%*
Bronx - 33.26%
Brooklyn - 36.25%
Queens - 42.91%
Staten Island - 62.265%

3) *(A subset of the result. Run the notebook "DOB_Permit_Issuance_2" to see the whole final data.)*



	BOROUGH	Street Name	Residential	Bldg Type	Zip Code	Surrounding Areas	Surrounding Ranks	Subway Nearby
1340	MANHATTAN	THIRD AVENUE	YES	2.0	10028	Upper East Side;	17;	Yes
3763	MANHATTAN	THIRD AVENUE	YES	2.0	10016	Garment District;Flatiron District;Gramercy Pa...	3;8;21;38;	Yes
3766	MANHATTAN	THIRD AVENUE	YES	2.0	10016	Garment District;Flatiron District;Gramercy Pa...	3;8;21;38;	Yes
4229	BROOKLYN	THIRD AVENUE	YES	2.0	11215	DUMBO;Boerum Hill;Carroll Gardens;Gowanus;Park...	5;7;15;16;23;26;34;36;49;	Yes
4501	MANHATTAN	THIRD AVENUE	YES	2.0	10016	Garment District;Flatiron District;Gramercy Pa...	3;8;21;38;	Yes

4) *445 days, 11 hours - Bronx*
437 days, 10 hours - Queens
460 days, 11 hours - Brooklyn
405 days, 10 hours - Manhattan
423 days, 10 hours – Brooklyn