DATA MINING: HOMEWORK 1
NAME: Meya Stephen Kenigbolo

QUESTION: Identify and present as a concise list various technical data analysis and visualisation methods used for the above research. Make first a list. And follow that with an extended list with an example sentence or illustration from that analysis.

1. **Maps**: Maps used in the analysis can be classified into three categories which include Road Maps, Animated maps and Physical maps
· *Road maps*: These were used to show major roads, some minor highways, airports etc. In real practice, People use road maps to plan trips and for driving directions.
· *Animated map*: These are maps which are usually in gif format
· *Physical map*: These are maps that show the physical features of an area including and not limited to features such as mountains, rivers, lakes etc. Water bodies are usually shown in blue while Green and orange are used to show the differences in the elevation of the land with Green indicating places with lower land elevation and orange indicating places with higher land elevations
2. **Graphs**: These also can be further classified into Bar Graph, Histogram, Line Graphs according to the way they were used to visualize the data
· *Bar Graph*: Graphs that use both words and numbers in the plot
· *Histogram*: Graphs where data is shown as numbers and in an order
· *Line Graph*: Shows points plotted on a graph and connected by lines
3. **Tables**: These are used to tabulate and compare data
4. **PostgreSQL**: Was used to store the data (NYC Taxi data).
5. **PostGIS**: Was used to perform geographic calculations, including the heavy lifting of mapping latitude/longitude coordinates to NYC census tracts and neighbourhoods (NYC Taxi data).

Road maps:

* Here's a map of Murray Hill, the most popular B&T destination, where each dot represents a single Saturday evening taxi trip originating at Penn Station:
* Drop offs for Saturday evening taxi rides originating at Penn Station
* New York City late Night travel index - 76% of the taxi pickups that occur in one of East Williamsburg's census tracts happen between 10 PM and 5 AM, the highest rate in the city. A paltry 5% of taxi pickups in some Upper East Side tracts occur in the late night hours

Line Graph:

* ***Line Graph of Brooklyn Monthly Taxi Pickup Graph***: Graph of taxi pickups in Brooklyn, the most populous borough, split by cab type. From 2009–2013, a period during which migration from Manhattan to Brooklyn generally increased, yellow taxis nearly doubled the number of pickups they made in Brooklyn.

* ***Line Graph of Uber vs Taxi Pickups in Brooklyn***: As of June 2015, the most recent data available when I wrote this, Uber accounts for more than twice as many pickups in Brooklyn compared to yellow taxis, and is rapidly approaching the popularity of green taxis

* ***Line Graph of Manhattan Monthly Taxi Pickups***: Manhattan, not surprisingly, accounts for by far the largest number of taxi pickups of any borough. In any given month, around 85% of all NYC taxi pickups occur in Manhattan, and most of those are made by yellow taxis. Even though green taxis are allowed to operate in upper Manhattan, they account for barely a fraction of yellow taxi activity:

* ***Line Graph of Uber vs Taxi pickups in Manhattan***: Uber has grown dramatically in Manhattan as well, notching a 275% increase in pickups from June 2014 to June 2015, while taxi pickups declined by 9% over the same period.

* ***Line Graph of Uber vs Taxi pickups in Queens***: Queens still has more yellow taxi pickups than green taxi pickups, but that's entirely because LaGuardia and JFK airports are both in Queens, and they are heavily served by yellow taxis. And although Uber has experienced nearly Brooklyn-like growth in Queens, it still lags behind yellow and green taxis, though again the yellow taxis are heavily influenced by airport pickups:

* ***Line Graph of Uber vs Taxi pickups at JFK and LaGuardia Airports***: If we restrict to pickups at LaGuardia and JFK Airports, we can see that Uber has grown to over 100,000 monthly pickups, but yellow cabs still shuttle over 80% of car-hailing airport passengers back into the city

* ***Line Graph of Northside Williamsburg Taxi Pickup***: The Northside neighbourhood is known for its nightlife: a full 72% of pickups occur during the late night hours.

* ***Graph for Cash vs Credit NYC Taxi Payment***: For example, did you know that in January 2009, just over 20% of taxi fares were paid with a credit card, but as of June 2015, that number has grown to over 60% of all fares?

* ***Graph for Cash vs Credit by Total Payments***: And for more expensive taxi trips, riders now pay via credit card more than 75% of the time:
* ***Graph for Travel time from Midtown, Manhattan to LaGuardia Airport***
* ***Graph for Travel time from Midtown, Manhattan to JFK Airport***
* ***Graph for Travel time from Midtown, Manhattan Newark Airport***

Physical map

* ***Google Maps*** estimates about an hour travel time on public transit from Bryant Park to JFK, so depending on the time of day and how close you are to a subway stop, your expected travel time might be better on public transit than in a cab, and you could save a bunch of money. The stories are similar for traveling to LaGuardia and Newark airports, and from other neighbourhoods. You can see the graphs for airport travel times from any neighbourhood by selecting it in the dropdown below:

* ***Map of Northside Williamsburg***: According to taxi activity, the most ascendant census tract in the entire city since 2009 lies on Williamsburg's north side, bounded by North 14th St to the north, Berry St to the east, North 7th St to the south, and the East River to the west

Animated Map

* ***Map of Taxi pickups in Northside Williamsburg***: Even before the boro taxi program began in August 2013, Northside Williamsburg experienced a dramatic increase in taxi activity, growing from a mere 500 monthly pickups in June 2009, to 10,000 in June 2013, and 25,000 by June 2015.
* ***Map of East Hampton's exclusive Further Lane***
* ***Map showing Hudson River Greenway and Goldman Sachs HQ***: Goldman Sachs lends itself nicely to analysis because its headquarters at 200 West Street has a dedicated driveway, marked "Hudson River Greenway" on this Google Map

Histogram

* ***Histogram of 72nd and Broadway Wall Street Taxi Travel Times***
* ***Histogram of Snowfall vs New York Daily Taxi Trips***
* ***Histogram of Precipitation vs New York Daily Taxi Trips***
* ***Histogram of Goldman Sachs weekday Taxi drop-offs at 200 West Street***
* ***Histogram of Citigroup weekday Taxi drop-offs at 388 Greenwich Street***

Tables

* Uber's ratio of 69% on 1/26/15 means that there were 69% as many Uber trips made that day compared to Uber's daily average from 1/19–1/25 on every single inclement weather day in 2015, in both rain and snow, Uber provided more trips relative to its previous week's average than taxis did

QUESTION: Identify the business value of potential data mining and analysis of the data of this type.
"There are investors out there who use satellite imagery to make investment decisions, e.g. if there are lots of cars in a department store's parking lots this holiday season, maybe it's time to buy. You might be able to do something similar with the taxi data: is airline market share shifting, based on traffic through JetBlue's terminal at JFK vs. Delta's terminal at LaGuardia? Is demand for lumber at all correlated to how many people are loading up on IKEA furniture in Red Hook?"
The above paragraph gotten from the passage illustrates the business value of the data. This gives credence to the proposition that proper analysis of data can result in a huge business value for

QUESTION (Cont.): Could you estimate the value in $$$ for Uber or NYC if some actions would follow this analysis?
Uber could increase their revenue by up to 1million dollars if they

QUESTION: A company makes computer discs. It tested a random sample of discs from a large batch and found that the probability of any disc being defective is 0,025. Bob buys two discs. Calculated the probability that

1. both discs are defective

Probability that one disk is defective = 0.025

Probability that both disks are defective = 0.025 * 0.025

=0.000625

2. That only one disc is defective.

D = Defective, G = Not Defective
Probability of defective disk (D) = 0.025
Probability of not defective disk = 1 – 0.025 = 0.975
Possible outcome that one is defective = [DG, GD]
Probability that one is defective = DG + GD = ([0.025 * 0.975] + [0.975 * 0.025]) = 0.04875

3. The company found 4 defective discs in the sample they tested. How many discs were likely tested?

P (D) = Probability of defective disk, n (t) = number of tested disks, n(s) = number of elements in Sample space.

P (D) = 0.025, n (t) = 4, n(s) =?

P (D) = n (t) / n(s)

Since we are looking for n(s), we will make n(s) the subject of the formula

N (s) = n (t) / P (D)

= 4 / 0.025 = 160

QUESTION: At the exam there is 0.8 probability that student has prepared and 0.2 that he has not prepared. Those who are prepared have 0.7 probability of success, those who have not prepared have 0.4 probability of success. What is the probability that randomly selected student will succeed?

Solution

We have two events from the above question
Event A = Situation that the student prepared
Event A~ = Situation that the student did not prepare
Event B = Situation where selected student succeeds
P(a) = Probability of Event A, P(a~) = Probability of event A~, P(b) = Probability of event B
P(a) = 0.8
P(a~) = 0.2
P(b|a) = 0.7
P(b|a~) = = 0,4
P(b ∩ a) = P(b|a)* P(a) = 0,7 * 0,8 = 0,56
P(b ∩ a~) = P(b|a~) *P(a~) = 0,4 * 0,2 = 0,08
P(b) = P(b ∩ a) + P(b ∩ a~) = 0,56 + 0,08 = 0,64