Final Project

Case Study in Mobile Analytics

Your Last Name, Your First Name

Deep Azure@McKesson
Dr. Zoran B. Djordjević

Problem Statement

- Two different businesses, a healthcare company and restaurant are looking to determine different information for their business planning based on using the same data set in Beijing, China of over 20 million records that contains GPS locations from mobile user data to answer the following meaningful insights noted below for Subway and United Healthcare. The technology used is R (for data wrangling and statistical analysis), JMP SMS (statisical analysis) and D3 for Visualization.
  - Restaurant Location
    - Subway, a chain restaurant company is looking for a new location for opening a restaurant in Beijing.

    Our analysis will determine:

    - Where are the most congested areas in the city, for lunch and dinner time?

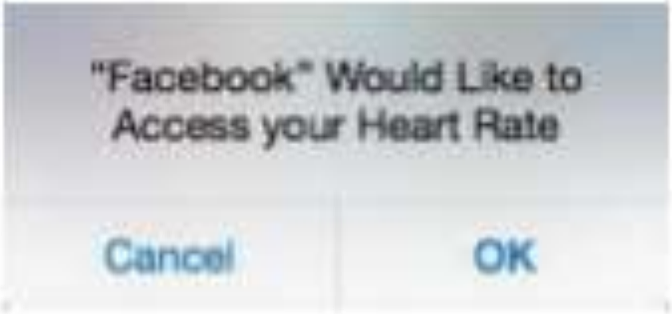
  - Healthiest Users
    - United Healthcare, a health insurance company is interested in the health behavior of its customers based on how they travel to work.

    Our analysis will determine:

    - Who are the healthiest users based on walking as the method of transportation?

Background

- Two-thirds of Americans are smartphone owners
- Millions of smartphones collect data and send this data “back home” to their companies
- The inspiration for this project came from an alert from the Facebook App on Apple Watch trying to get access to Heart Rate of its user



"Facebook" Would Like to Access your Heart Rate

- High tech companies are essentially using their users as human sensors to collect information
- Security and Privacy issues aside, vast amount of data is being collected to provide an incredible opportunity for various research and development.

Data Set

- Data can be downloaded from Microsoft Research: http://ftp.research.microsoft.com/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/Geolife%20Trajectories%201.3.zip?

Microsoft Research

- The unzipped and combined data set is 1.5GB with 23 million rows and 7 columns
- The dataset includes location of users in longitude and latitude, the timestamp, the method of transportation, and altitude.
- It includes a broad range of users' outdoor movements such as going home and going to work, sports activities, shopping, etc.
- Data from cities of China, USA, and Europe. However, the majority of the data was created in Beijing, China.
- Dataset came in thousands of PLT files for each trajectory recorded. Also these PLT files were all in separate folders for each user.
- Each PLT file had 6 lines of useless information

Technology and Software

- Cygwin is used to run UNIX commands. These UNIX commands are used mostly for wrangling the data. Cygwin can be installed from: https://cygwin.com/install.html

- JMP is statistical and graphing software. SAS JMP is available for trial and to buy: http://www.jmp.com/en_us/offers/free-trial.html

- D3 is a JavaScript library used for Data Visualization.

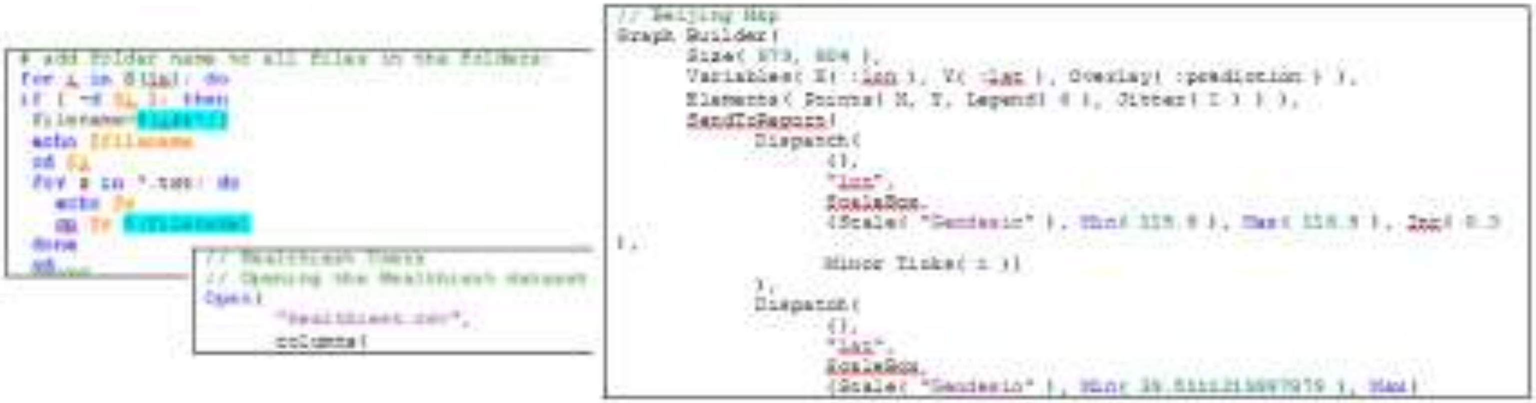
- R Studio is also utilized for statistical analysis and generating the heat maps. Can be downloaded free: http://www.rstudio.com/products/rstudio/download/

My Hardware Environment

- Windows Operating System version 7
- Linux, Red Hat version 6

Code Overview

- Overview of Steps:
  - Install R and JMP software
  - Data wrangling: combining the data, cleaning the data, formatting the data
  - Exploratory Data Analysis in JMP and R
  - Formatting, sub setting, heat maps in R
  - Visualization in JMP with pie charts and histogram
- Sample of code to read data and build graphs after our analysis is performed.



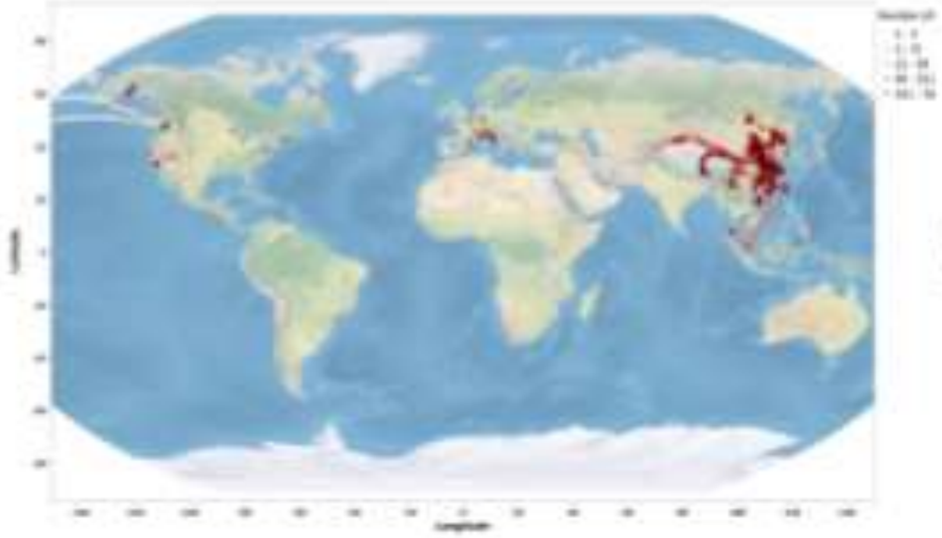
```
# add folder name to all files in the folders
for i in $(ls); do
if [ -d $i ]; then
 filename=[illegible]
 echo $filename
 cd $i
 for x in *.txt; do
   echo $x
   mv $x [illegible]
 done
 cd..
```

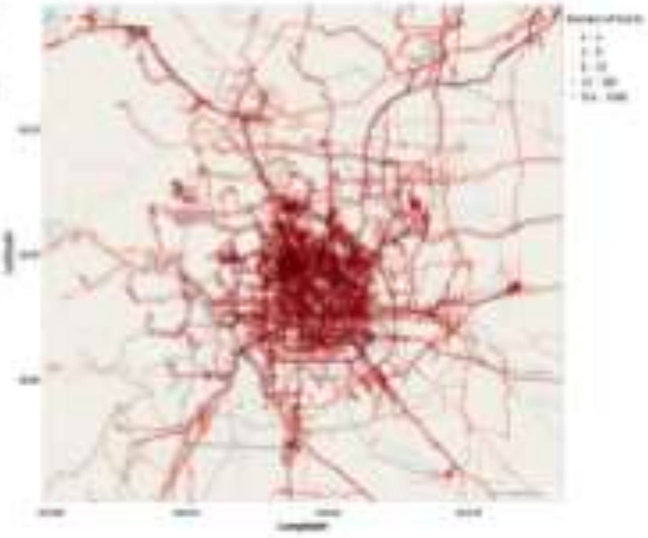
```
// Healthcash Users
// Opening the Healthcash dataset
Open(
	"healthcash.csv",
	columns(
```

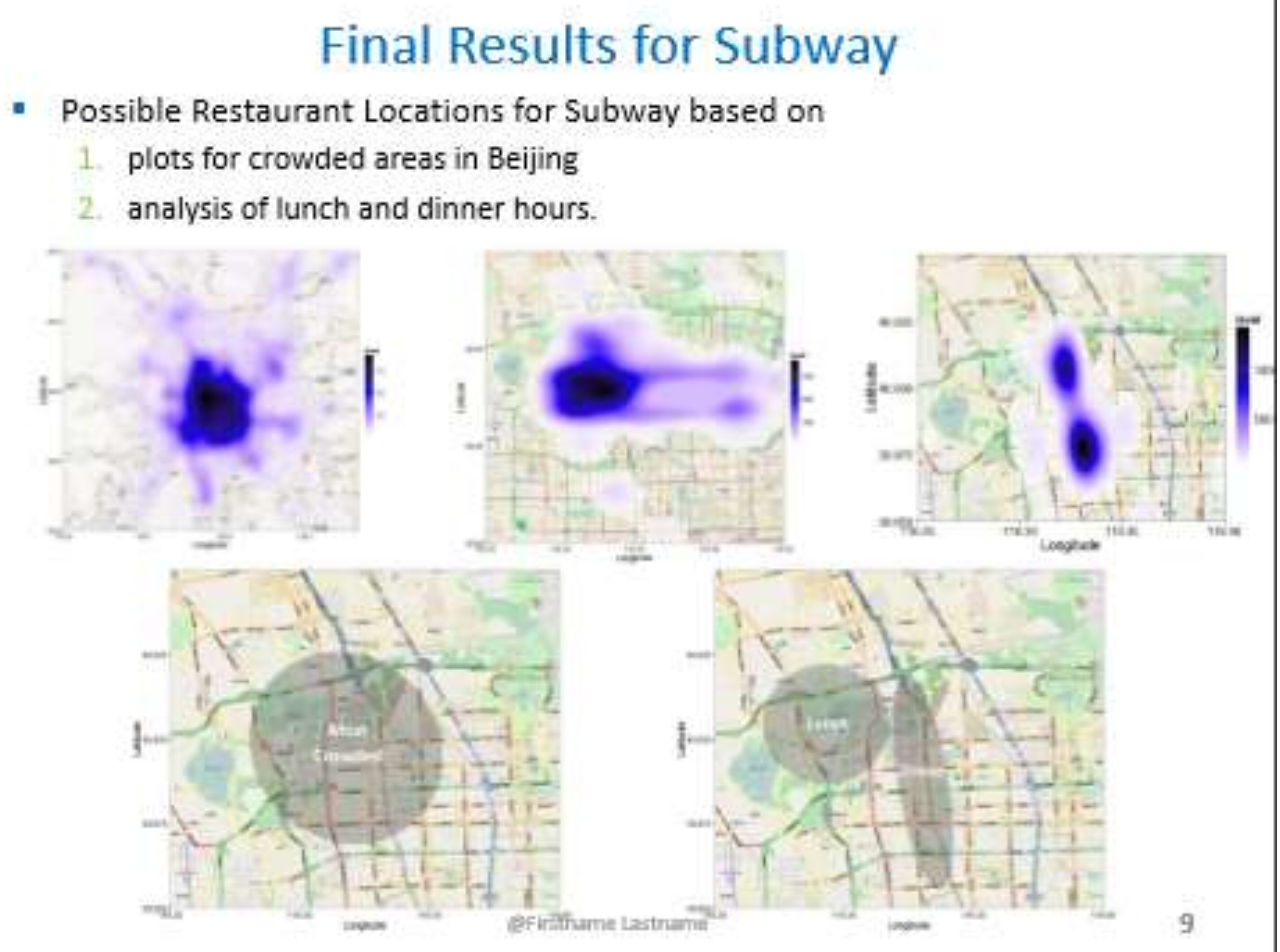
```
// Beijing Map
Graph Builder(
	Size( 573, 504 ),
	Variables( X( :Lon ), Y( :Lat ), Overlay( :prediction ) ),
	Elements( Points( X, Y, Legend( 6 ), Jitter( 1 ) ) ),
	SendToReport(
		Dispatch(
			{},
			"Lon",
			ScaleBox,
			{Scale( "Geodesic" ), Min( 115.9 ), Max( 116.5 ), Inc( 0.2
),
			Minor Ticks( 1 )}
		),
		Dispatch(
			{},
			"Lat",
			ScaleBox,
			{Scale( "Geodesic" ), Min( 39.8111213997979 ), Max(
```

Preliminary Results

- Preliminary Exploratory Data Analysis is done to get a feel where the data is mostly located
- After cleaning the data, data is plotted on maps
- Plots created in JMP show GPS locations of many cities around the world, and Beijing





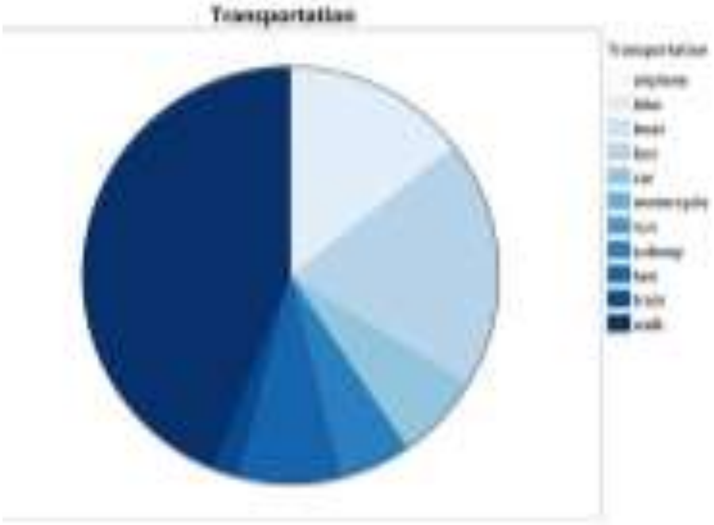


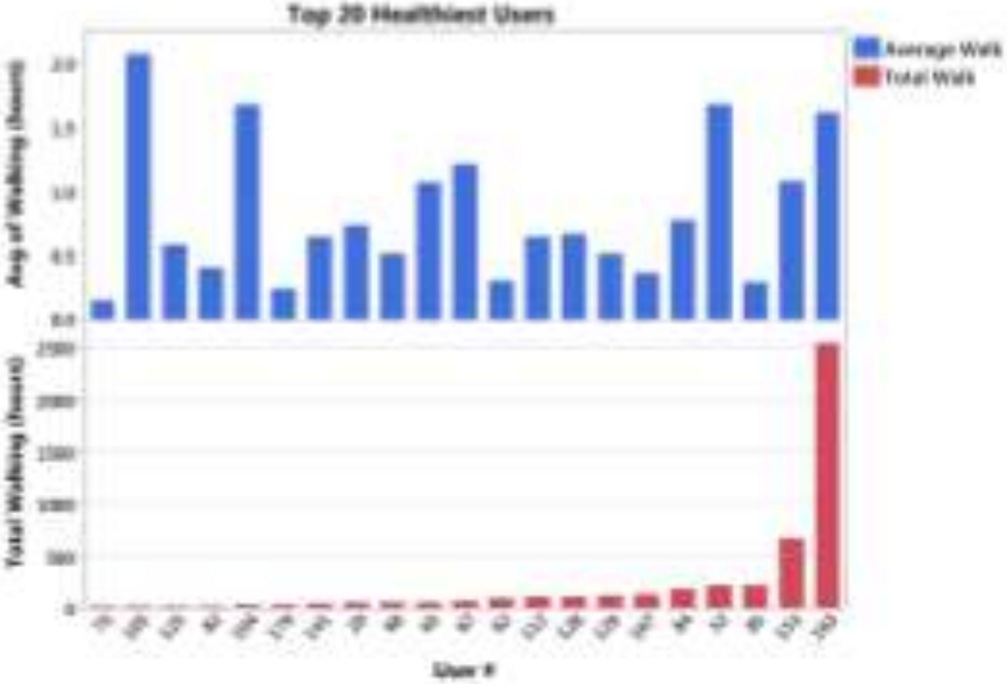
Final Results for Subway

- Possible Restaurant Locations for Subway based on
  1. plots for crowded areas in Beijing
  2. analysis of lunch and dinner hours.

Final Results for United HealthCare

- Healthiest Users Analysis for the United HealthCare.
  - Transportation used by mobile users are noted in the left .
  - On the right the healthiest is determine by the amount of walking.





Lessons Learned

- In some instances, R gave few inconsistent errors handling a 23 Million row dataset.
- This was apparent specially in creating heat maps, which R was a bit slow and troublesome. Often some points wouldn't get plotted, and R would give major fail errors.
- The dataset was distributed in thousands of files and wrangling the data took long time
- The data set included millions of GPS locations, but from only hundreds of users.
- Dataset only has few hundreds of users
- It was determined that the most crowded location in Beijing was Tsinghua University! Most likely caused by the fact that most of the users were researchers and students that traveled to the university often.
- Nevertheless, the analysis and codes still hold for another big data set that includes thousands of random users, to draw more meaningful results.

Future Work

- R on HDFS and Hadoop would have been much faster
- The RImpala package is provided for Cloudera Impala to integrate with R and HDFS to achieve faster and more reliable results.
- The documentations for RImpala can be found at:
  - http://blog.cloudera.com/blog/2013/12/how-to-do-statistical-analysis-with-impala-and-r/
  - http://www.cloudera.com/content/cloudera/en/documentation/cloudera-impala/v2-0-x/topics/impala_noncm_installation.html

- On the other hand, the benefit of doing analysis this way was that installing more software packages and potential problems would be avoided, especially in the given short period of time.

- Additional metadata within the dataset could draw other meaningful results.

YouTube URLs, GitHub URL, Last Page

- Two minute (short):
- 15 minutes (long):
- GitHub Repository with all artifacts: