**GBA 6050 Group Project Report**

**Introduction**

There are several forms of transportation nowadays. With busy streets full of cars and pedestrians, we can gather sufficient data on means of transportation for civilian use. In today’s state, it has become no surprise how pedestrians have shifted their automobile preferences to biking, walking, or taking the bus. When determining what means of transportation to partake in, several parameters must be taken into account. A few roleplaying factors revolve around time of the day, time, location, number of employed individuals, reasons for traveling, etc. When regarding the number of automobile drivers, bikers, bus takers, taxi riders, etc there is a sense of correlation that links all forms of transportation together. In California we are aware that the majority of people travel through vehicles due to distances and locations. However, the question at issue is determining what relationships belong between pedestrian trips and biking trips in California metropolitan and rural areas. As mentioned priorly, citizens have explored new means of transportation due to factors such as high gas prices.

**Description of Data Preparations**

Before attempting to create a relationship between pedestrian trips and biking and how they compare across metropolitan and rural areas, we decided to redefine the data and focus on a data set with far fewer factors. All information presented was valuable but would oversaturate our result and be misleading. There are several data sets that revolve around the same topic, for example, time or work, vehicle related sections, etc. Therefore, we decided to go with an approach that would analyze a broader side of all these data sets. Our main focus was to narrow down the data so we could emphasize data on metropolitan and rural areas. We inferred that pedestrians in metropolitan areas are more likely to bike or walk rather than travel through vehicles. As stated before, several parameters are taken into account by civilians to determine their forms of travel.

In order to make the data legible we had to go through various preprocessing and cleaning phases before we were able to consolidate on the relevant factors and split the data according to the rural/urban dichotomy. Once that was done and null values were eliminated, we rebalanced the data so that we could compare two equally sized representative samples across which we would draw comparisons. Using the split datasets and the combined dataset we were able to calculate several patterns using some simple statistics in Excel & Tableau.

**Excel Data Analysis / Story**

First, by taking the variables that related to vehicle availability per household we were able to notice some minor differences between our Urban and Rural data samples. For instance when looking at the relationship between the number of Household Vehicles and the number of Drivers in Household or Adults in Household, there is a closer correlation in Urban households, indicating that those are closer to having a vehicle per driver. Furthermore Household Vehicles and Household members which includes non-drivers shows a significantly stronger relationship in Urban households than in Rural households. This indicates that on average Urban households have less vehicles and less people that can drive them. Interestingly, this actually seems to demonstrate that more time is spent driving in rural households, despite more constrained resources.

Secondly, by shifting our attention to the variables representing the rates of use for the various methods of transportation, we were able to generate a single-factor ANOVA test (Analysis of Variance) for both the rural and the urban data samples. This test was used to track the significance of each variable within the tested data samples. With this approach we were able to see that there was a significant increase in the choice of walking among urban households than rural households. Although the difference in bicycle use was less pronounced, there is a noticeable increase across the board in the use of all modes of transportation, including bus train and taxi), except for Car and Rail which both showed a higher average and lower variance in rural households when compared to urban households. This seems to indicate that although there is a small increase in bicycle use in urban households, there is also a much greater walking rate as well. The only categories in which rural households have a higher utility rate are Rail and Car. All other categories reflect a greater usage in urban areas, which seems to lend some support to our initial hypothesis that there is a greater use of bicycles in Urban areas because the population is on average more active in these areas.

Finally, another set of variables that we selected for analysis was the use rate for various means of accessing the internet. We hoped this would shed some light on the use of internet technology in the two groups and if that lined up with our hypothesis pertaining to regional rates of activity. Interestingly enough our results actually showed a greater use of technology across all categories in urban households on average than in rural households. This is further supported by the negative skewness on urban variables which indicates that on a normal bell-curve distribution, the data is more dense at the higher end of the distribution, suggesting a majority of the sample values are weighted toward the higher end of its scope.

**Conclusion**

After extracting, prepping, and analyzing data, we concluded that there is a higher number of people walking and biking in urban areas; demonstrating higher rates of activity in most categories. There are fewer people per households located in rural areas that own cars or can drive cars. Nevertheless, civilians located in rural areas tend to drive cars more than walking or biking. In addition, we aimed to factor in additional parameters to calculate rates of activities of regions, Some variables included were internet usage in urban and rural areas. As a result, we determined that internet usage rates are relatively higher in urban households than in rural households. Internet usage was selected as a parameter to help us understand how active each region is. However in the end it seemed that our first prediction about urban areas being more active was correct, however, regarding internet usage we found no significant impact on travel or activity as the same group that showed higher activity was also found to use the internet more often. Therefore it does not follow that internet use is necessarily an indicator of a more sedentary lifestyle.

**Appendix**

**Word Document**

Introduction-Wihtman

Data Preparations- Norberto, Freddy, Ramandeep, Wihtman

Data Analysis-Norberto, Freddy, Ramandeep, Wihtman

Data Analysis/story- Norberto, Freddy

Conclusion- Ramandeep, Wihtman

**Powerpoint Presentation**

Introduction slides (1-4) -Wihtman

Step 1 (5-7)-Norberto

Step 2 (8-10)-Freddy

Step 3 (11-14)-Ramandeep

Q&A (14)-Wihtman