DATA 606 Data Project Proposal

Adam Gersowitz

Data Preparation

```
library(nasaweather)
library(sqldf)
nrow(atmos)
length(atmos)
head(atmos,10)

# I will create a dataframe of the information that could impact the temperature as well as the progres

temp <- atmos[c(1:6)]

# I will also create dataframes that store average information on temperature by date and location
temp<-sqldf("select a.*, lat||'-'||long ll from temp a")

meantemp_coord <-sqldf("select avg(temp) mean_temp,avg(surftemp) mean_surftemp, lat, long from temp grou

meantemp_date  <-sqldf("select avg(temp) mean_temp,avg(surftemp) mean_surftemp, month, year from temp gr

meantemp_year  <-sqldf("select avg(temp) mean_temp,avg(surftemp) mean_surftemp,  year from temp group by

meantemp_coordyear  <-sqldf("select avg(temp) mean_temp,avg(surftemp) mean_surftemp, lat,long,year,ll fr

#Lastly I'll create a dataframe that captures the percent change from month to month

temp_percent_change  <-sqldf(" select tm.lat, tm.long, tm.month, tm.year, tm.temp last_month_temp, tm.su
                               (surftemp-last_month_surftemp)/last_month_surftemp*100 as chng_surftemp
                               from
                              (select temp,surftemp, month, year, lat, long
                              from temp) tm
                             join
                             (select temp as last_month_temp, surftemp as last_month_surftemp, lat, long
                            case when month = 12 then year-1 else year end year,
                            case when month = 12 then 1 else month+1 end month from temp) lm on
                             lm.lat = tm.lat and lm.long = tm.long and lm.year = tm.year and lm.month =
                             order by  tm.lat, tm.long

                             ")

head(meantemp_coordyear,10)
```

Research question

Has the average monthly temperature changed at the same rate across different coordinates in Central America?

Cases

Each case represents an atmospheric weather reading of coordinates in Central America. There 41,472 cases in the data set.

Data collection

The data was collected by NASA and is data from the 2006 ASA data expo.

Type of study

This is an observational study.

Data Source

The data can be found at http://stat-computing.org/dataexpo/2006/. For this project the data was pulled from the nasaweather package. (https://blog.rstudio.com/2014/07/23/new-data-packages/)

Dependent Variable

The response variables are the temperature and surface temperatures which are numerical.

Independent Variable

Teh explanatory variable is the cooridnates (laititude and longitude) which is categorical.

Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

```
summary(temp_percent_change$chng_temp)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -7.430856 -0.303132  0.000000  0.001669  0.304054  7.601291
```

```
#install.packages("Hmisc")

summary(temp_percent_change$chng_surftemp)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -6.483626 -0.337838  0.000000  0.002964  0.342818  4.979947
```

```
library(ggplot2)
ggplot(data = meantemp_year , aes(x = year, y = mean_temp)) +
     geom_line() +
     geom_point() +
     scale_x_discrete(breaks = meantemp_year$year, labels = meantemp_year$year)
```