DSC520 Assignment 10.3 Final Project Step 3
Name: Ruth Maina
Date: 3/4/2023

Introduction:

The topic of obesity is one that has continued to be of high interest here in America. Back in the 70s, the obesity rate in America used to be in the teens. In a span of about 50 years, the rate has continued to steadily increase, and it is projected to reach 50 percent by the year 2030.

The problem statement to be addressed:

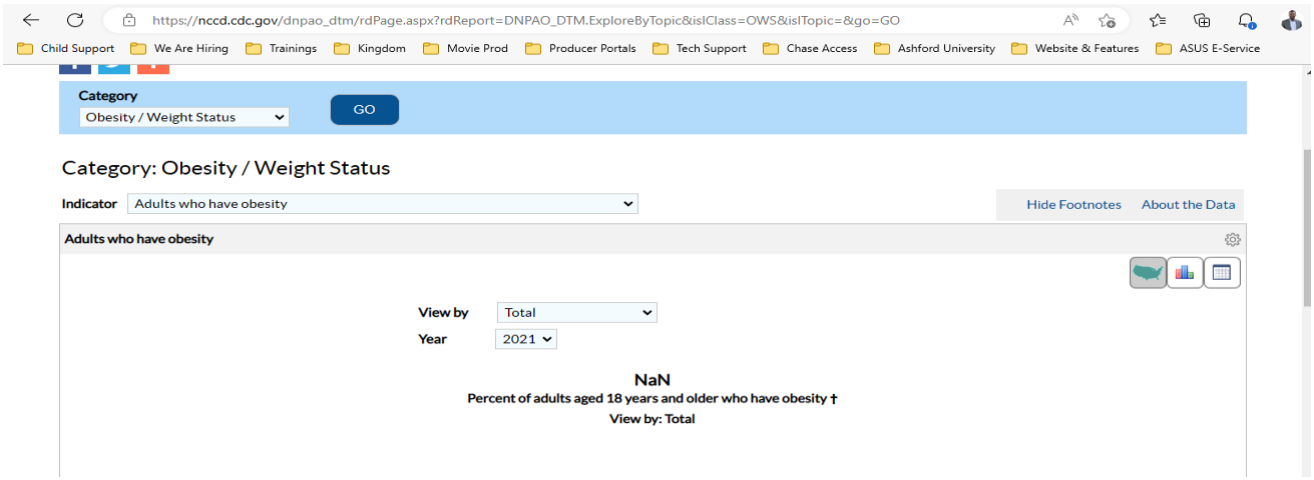
My main data point is to explore obesity data for the last ten years, so as to let the data tell a story on how this is trending, and also perform a prediction of where this could go in the next few years. I'm also curious to see if the growth prediction value of 50 percent by 2030 is accurate.

Data preparation steps:

- ✓ The obesity data is available in various formats from the source such as pdf, excel, csv. I chose to use csv format because its familiar and straightforward.
- ✓ I downloaded the csv file and reviewed the file, scrutinizing the available data fields by copying-pasting the column titles in transposed format so as to get a good view of the field names
- ✓ I noticed that the Percent of adults aged 18 years and older who have obesity data is available and downloadable by year, from 2011 to 2021; Also, it is available by Education, Gender, Income, and Race/Ethnicity, which is very exciting to see. I chose to focus on adults

age group even though there's other age groups available such as adolescents, two- four-year-olds, as well as 3–23-month-olds.

- ✓ I merged the by year data and saved the files from 2011 to 2021, in the same data directory I have been using for the course. I noticed my different data sets we're presenting the same data, just in different groupings so I merged them all into one dataset



Data cleansing steps in R:

- ✓ Loaded the csv data into R, into a data frame.

```
> obesityData_df <- read.csv("data/adults_over18.csv")
> head(obesityData_df)
```

- ✓ Reviewed the number of observations was 602 and 43 variables, which is a good size to work with. Of all these variables, I have selected to work with just the obesity percentage field. It comes in 3 fields, an actual value, a high confidence limit and a low confidence limit. I will work only with the actual value.
- ✓ I will need only one variable which contains obesity percentage values. I checked the variable data type and noticed its stored as 'chr'. I converted it to int:

```
> is.integer(obesityData_df$Data_Value)
[1] FALSE
> obesityData_df$Data_Value <- as.integer(obesityData_df$Data_
Value)
```

```
> is.integer(obesityData_df$Data_Value)
[1] TRUE
```

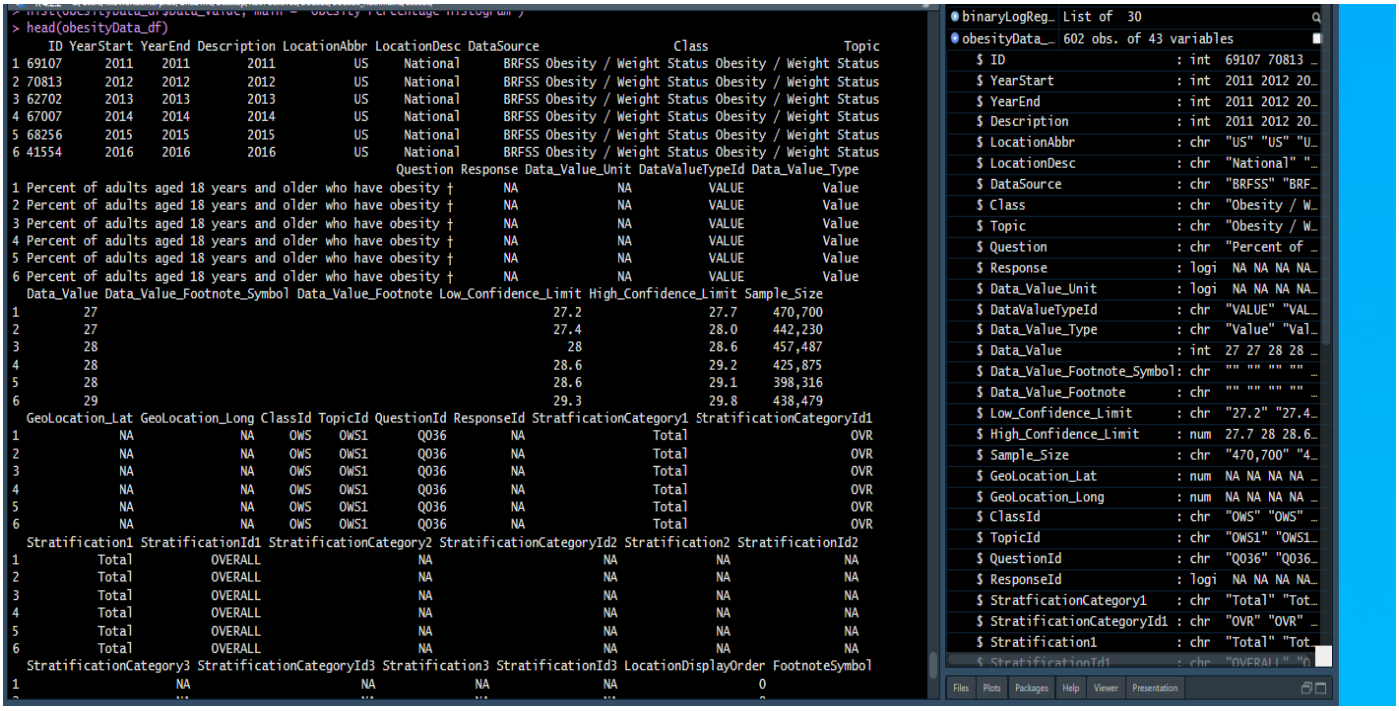
✓ Using summary stats, I received below output which revealed some missing values:

```
> describe(obesityData_df$Data_Value)
obesityData_df$Data_Value
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     589       13       21    0.995    29.77    4.609     23.0     24.8     27.0     30.0     33.0     35.0     36.0

lowest : 20 21 22 23 24, highest: 36 37 38 39 40
>
```

✓ A few values in my variable contain a character value of '~' which is an irregularity. I checked these after conversion to integer and notice that they're NA. I will omit NA in my calculations using na.action=na.omit

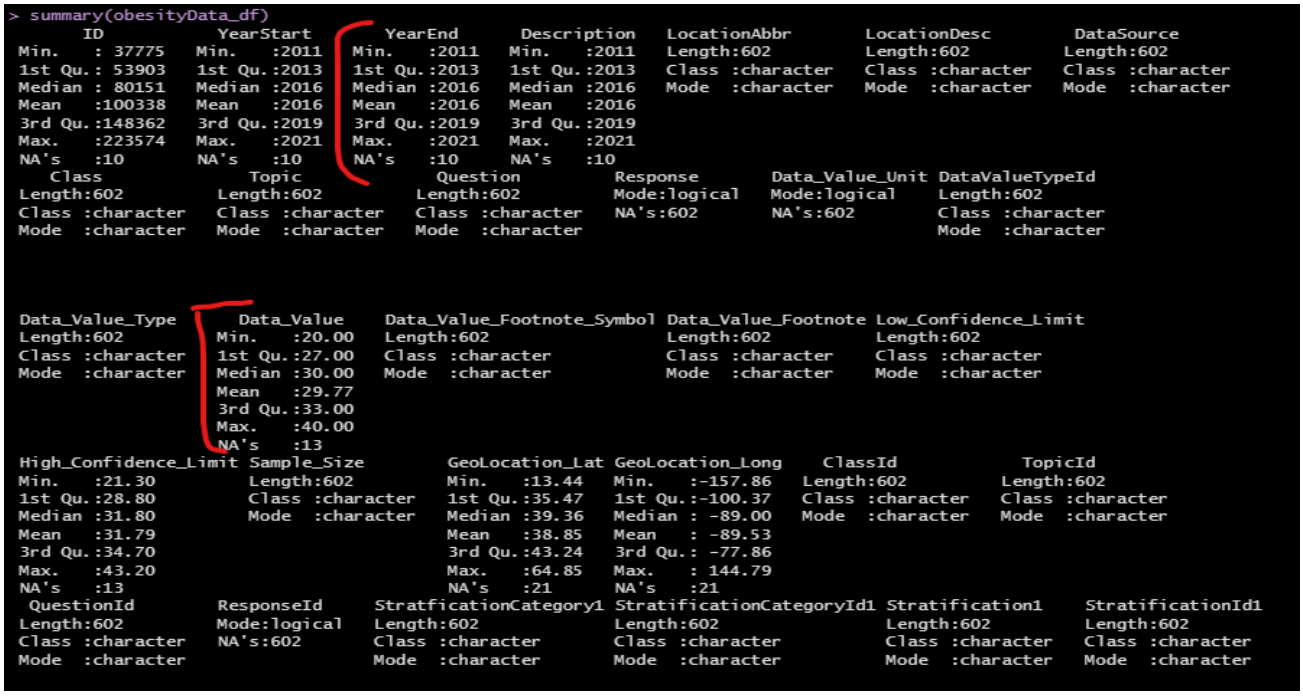
```
> is.na(obesityData_df$Data_Value)
```

Final dataset:



Analysis steps and Insights:

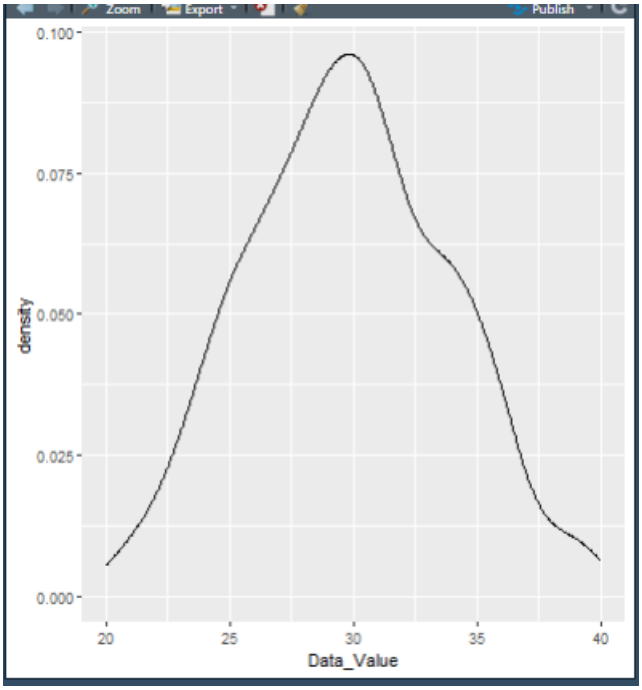
- A review of summary statistics indicates obesity percentages steadily increasing every single year



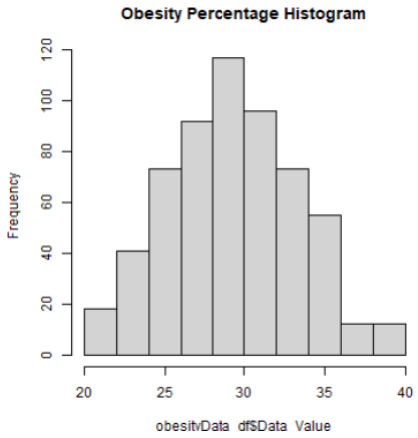
```
> summary(obesityData_df)
       ID            YearStart       YearEnd       Description    LocationAbbr       LocationDesc        DataSource
 Min.   : 37775   Min.   :2011   Min.   :2011   Min.   :2011   Length:602         Length:602         Length:602
 1st Qu.: 53903   1st Qu.:2013   1st Qu.:2013   1st Qu.:2013   Class :character   Class :character   Class :character
 Median : 80151   Median :2016   Median :2016   Median :2016   Mode  :character   Mode  :character   Mode  :character
 Mean   :100338   Mean   :2016   Mean   :2016   Mean   :2016
 3rd Qu.:148362   3rd Qu.:2019   3rd Qu.:2019   3rd Qu.:2019
 Max.   :223574   Max.   :2021   Max.   :2021   Max.   :2021
 NA's   :10       NA's   :10     NA's   :10     NA's   :10
    Class              Topic            Question          Response       Data_Value_Unit DataValueTypeId
 Length:602         Length:602         Length:602         Mode:logical   Mode:logical    Length:602
 Class :character   Class :character   Class :character   NA's:602       NA's:602        Class :character
 Mode  :character   Mode  :character   Mode  :character                                  Mode  :character



 Data_Value_Type     Data_Value     Data_Value_Footnote_Symbol Data_Value_Footnote Low_Confidence_Limit
 Length:602         Min.   :20.00   Length:602                 Length:602          Length:602
 Class :character   1st Qu.:27.00   Class :character           Class :character    Class :character
 Mode  :character   Median :30.00   Mode  :character           Mode  :character    Mode  :character
                    Mean   :29.77
                    3rd Qu.:33.00
                    Max.   :40.00
                    NA's   :13
 High_Confidence_Limit Sample_Size        GeoLocation_Lat GeoLocation_Long    ClassId            TopicId
 Min.   :21.30         Length:602         Min.   :13.44   Min.   :-157.86   Length:602         Length:602
 1st Qu.:28.80         Class :character   1st Qu.:35.47   1st Qu.:-100.37   Class :character   Class :character
 Median :31.80         Mode  :character   Median :39.36   Median : -89.00   Mode  :character   Mode  :character
 Mean   :31.79                            Mean   :38.85   Mean   : -89.53
 3rd Qu.:34.70                            3rd Qu.:43.24   3rd Qu.: -77.86
 Max.   :43.20                            Max.   :64.85   Max.   : 144.79
 NA's   :13                               NA's   :21      NA's   :21
  QuestionId         ResponseId     StratficationCategory1 StratificationCategoryId1 Stratification1    StratificationId1
 Length:602         Mode:logical   Length:602             Length:602                Length:602         Length:602
 Class :character   NA's:602       Class :character       Class :character          Class :character   Class :character
 Mode  :character                  Mode  :character       Mode  :character          Mode  :character   Mode  :character
```

- To get a feeling of the data distribution, I used the below density plot, histogram, and scatter plots to review the data. The distribution of obesity is mostly normal/symmetrical and unimodal with one clear peak in the data, without outliers or skew. This tells me that a normal distribution could accurately be used as a model for obesity data.
- In addition to viewing Obesity by Year from 2011-2021, I also have a scatterplot view by State which reveals three states with the highest percentage being KY, MS, and WV respectively, all around 40% value.

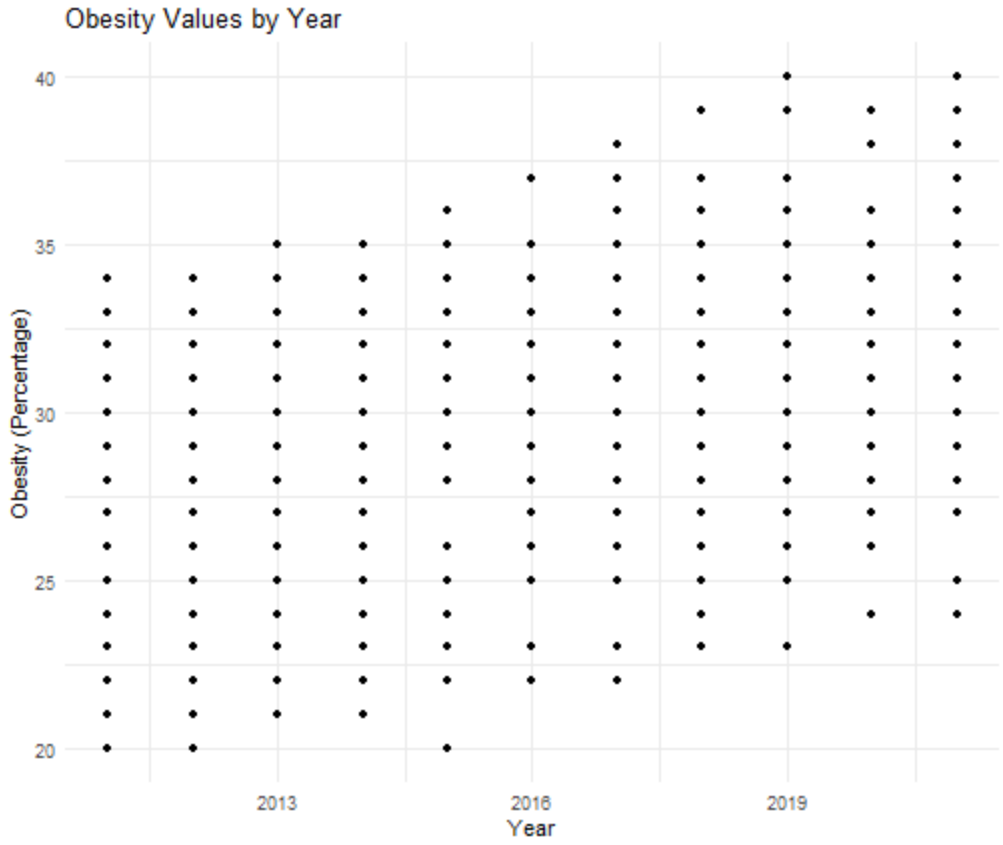
```
> ggplot(obesityData_df, aes(x=Data_Value)) + geom_density()
```



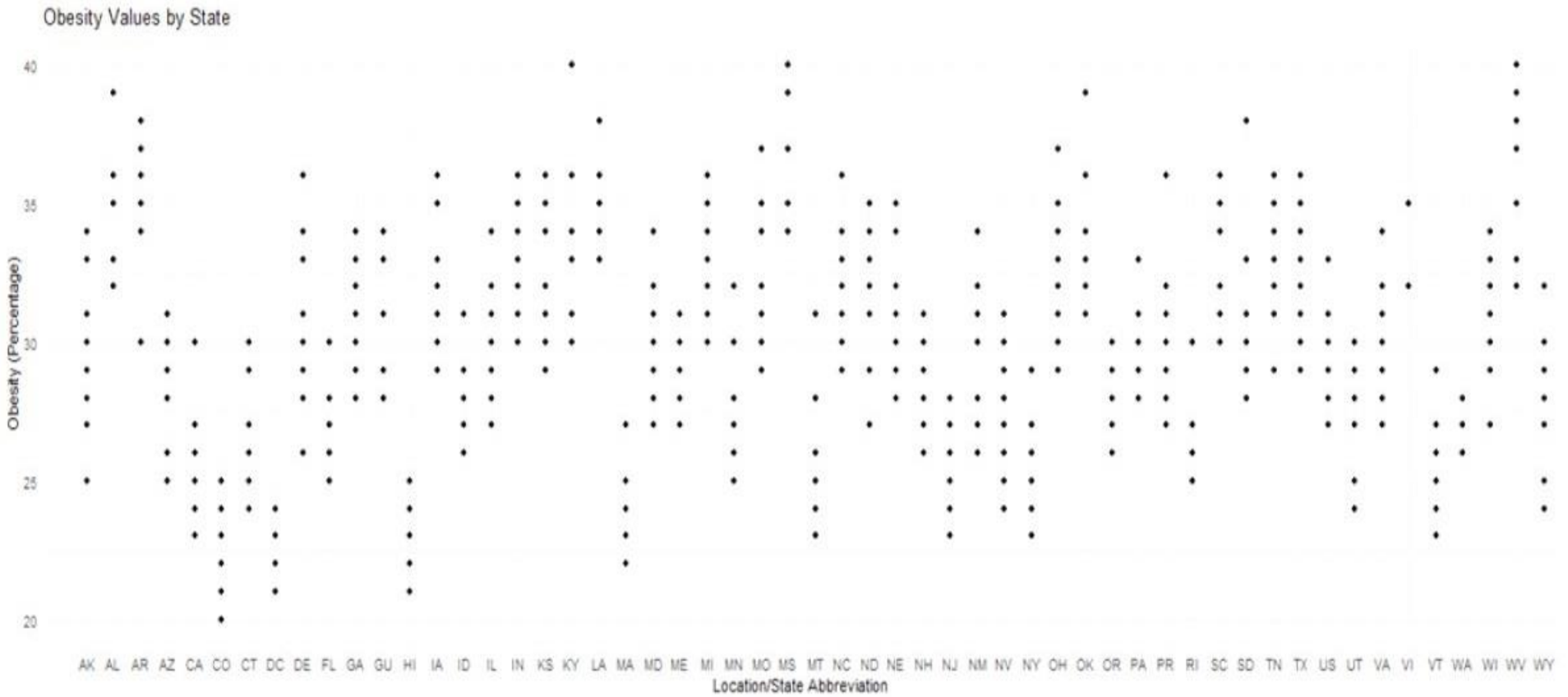

```
hist(obesityData_df$Data_Value, main = "Obesity Percentage His
togram")
```




```
> ggplot(obesityData_df, aes(x=YearStart, y=Data_Value)) + geom_poi
nt() +
+     ggtitle("Obesity Values by Year") + xlab("Year") + ylab("Obes
ity (Percentage)")
```




```
> ggplot(obesityData_df, aes(x=LocationAbbr, y=Data_Value)) + geom_
point() +
+     ggtitle("Obesity Values by State") + xlab("Location/State Abb
reviation") + ylab("Obesity (Percentage)")
```


Obesity Values by State
Obesity (Percentage)
40
35
30
25
20
AK AL AR AZ CA CO CT DC DE FL GA GU HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS MT NC ND NE NH NJ NM NV NY OH OK OR PA PR RI SC SD TN TX US UT VA VI VT WA WI WV WY
Location/State Abbreviation

- Since I have 3 main variables, namely Year, State, and Obesity percentage/Data Value, I fit a linear model below using the Year variable as the predictor and Obesity as the outcome.

```
> obesity_lm <- lm(formula = Data_Value ~ YearStart, data=obesityData_df)
> summary(obesity_lm)

Call:
lm(formula = Data_Value ~ YearStart, data = obesityData_df)

Residuals:
    Min      1Q  Median      3Q     Max
-9.1784 -2.4644  0.1075  2.5356  8.5356

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.122e+03  9.590e+01  -11.70   <2e-16 ***
YearStart    5.715e-01  4.757e-02   12.01   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.635 on 587 degrees of freedom
  (13 observations deleted due to missingness)
Multiple R-squared:  0.1974, Adjusted R-squared:  0.196
F-statistic: 144.3 on 1 and 587 DF,  p-value: < 2.2e-16
> obesity_lm

Call:
lm(formula = Data_Value ~ YearStart, data = obesityData_df)

Coefficients:
(Intercept)    YearStart
 -1122.3737       0.5715
```

- For below I am attempting to predict obesity for the next 8 years. The prediction indicates a steady percentage no more than 40 % , I am surprised the number has not increased gradually but its definitely levelled which could be accurate. This

```
> next_ten_years <- data.frame(YearStart = c(2022,2023,2024,2025,2026,2027
> linear_model <- lm(Data_Value ~ YearStart, data=obesityData_df)
> predict(linear_model, newdata = next_ten_years)
       1        2        3        4        5        6        7        8
33.17888 33.75037 34.32186 34.89335 35.46484 36.03633 36.60782 37.17931 37
> predict(linear_model, newdata = next_ten_years, interval = 'confidence')
       fit      lwr      upr
```

```
1 33.17888 32.54836 33.80939
2 33.75037 33.03589 34.46484
3 34.32186 33.52132 35.12240
4 34.89335 34.00526 35.78144
5 35.46484 34.48811 36.44157
6 36.03633 34.97014 37.10252
7 36.60782 35.45154 37.76409
8 37.17931 35.93246 38.42616
9 37.75080 36.41298 39.08862

>
```

Implications:

- The key implication from the analysis is to prove that obesity is a weighty subject that should be tackled collectively by all
- The trend percentages for the last decade and future prediction indicate an increased disparity from 2011 – 2030, which would help the audience see the importance of the matter.
- While the last decade/actual data shows a sharp increase, the latter decade prediction shows a levelling off, which could mean that some causative factor could be helping avoid gradual increases compared to the last decade.

Limitations:

- The model accuracy could be increased by adding additional years of data. For example, instead of starting at 2011. We could go back several more years so as to fine tune the prediction.

Concluding Remarks:

- The above analysis has shed light on obesity, providing analysis of the past decade and a potential prediction of the next ten years. The obesity issue can be tackled in many ways

and intervention opportunities could be explored from all angles, particularly from governing bodies and community education.

DATASET REFERENCES:

Behavioral Risk Factor Surveillance System (2020, November 10). Data Catalog. https://data.cdc.gov/api/views/hn4x-zwk7/rows.csv?accessType=DOWNLOAD

CDC (n.d.) Nutrition, Physical Activity, and Obesity: Data, Trends and Maps. https://nccd.cdc.gov/dnpao_dtm/rdPage.aspx?rdReport=DNPAO_DTM.ExploreByTopic&islClass=OWS&islTopic=&go=GO

National Center for Health Statistics(n.d). Health, United States. https://www.cdc.gov/nchs/hus/data-finder.htm.

Global Health Data Exchange (n.d). United States Physical Activity and Obesity Prevalence by County 2001-2011. https://ghdx.healthdata.org/sites/default/files/record-attached-files/IHME_USA_OBESITY_PHYSICAL_ACTIVITY_2001_2011.csv

OTHER GENERAL RESEARCH SUPPLEMENTAL REFERENCES:

CDC (n.d.). Causes of Obesity https://www.cdc.gov/obesity/basics/causes.html

CDC (n.d.). Adult Obesity Facts https://www.cdc.gov/obesity/data/adult.html

World Health Organization. (2021, June 9). Obesity and overweight. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

National Geographic. (2017, April 6). 5 “Blue Zones” Where the World’s Healthiest People Live https://www.nationalgeographic.com/books/article/5-blue-zones-where-the-worlds-healthiest-people-live

Dr.Mandal, A MD. (2019, February 27). Obesity and Fast Food. https://www.news-medical.net/health/Obesity-and-Fast-Food.aspx

NCHS Data Brief No. 322 (2018, October). Fast Food Consumption Among Adults in the United States, 2013–2016. https://www.cdc.gov/nchs/products/databriefs/db322.htm#:~:text=In%202013%E2%80%932016%2C%2036.6%25,adults%20aged%2060%20and%20over.

Wikipedia (n.d.). Obesity in the United States. https://en.wikipedia.org/wiki/Obesity_in_the_United_States#Contributing_factors