Ryan Long

11/11/2019

Overall, write a coherent narrative that tells a story with the data as you complete this section.

Summarize the problem statement you addressed.

Two contributing factors related to maintaining body weight are physical activity and eating nutritious foods. The question is to what degree does individual choices regarding physical activity and quality of nutrition aid the effort to prevent obesity?

A data set on Nutrition, Physical Activity and Obesity created by the Centers for Diease Control and Prevention was obtained from HealthData.gov. The raw data set was cleaned and prepared to contain the following variables for 2017 survey by State/Territory:

Obese: Percentage of Adults aged 18 years and older who have obesity.
NoActivity: Percentage of Adults who engage in no leisure-time physical activity.
Fruit1dy: Percentage of Adults consuming fruit less than one time daily.
Veg1dy: Percentage of Adults consuming fruit less than one time daily.

The following were removed for the following reasons Puerto Rico (Outlier), Virgin Island (incomplete), and US Composite (non-represntative).

The below depicts the data cleaning steps.

```
# Data Source: https://healthdata.gov/dataset/nutrition-physical-activity-and-obesity-behavioral-risk-factor-surveillance-system

#Import the raw data.
data<- read.csv("Nutrition__Physical_Activity__and_Obesity_-_Behavioral_Risk_Factor_Surveillance_System.csv")

#Filter criteria to be applied to the data. The first filter will target the attributes related to obesity, activity, and nutrition. The second filter will include the most recent survey data. Prior years did not include nutrition data. The third filter will exclude all category strata and will only pull the total.

qfilt<-c("Percent of adults who engage in no leisure-time physical activity", "Percent of adults aged 18 years and older who have an overweight classification", "Percent of adults aged 18 years and older who have obesity", "Percent of adults who report consuming fruit less than one time daily", "Percent of adults who report consuming vegetables less than one time daily")
yfilt<- "2017"
tfilt<- "Total"

#Apply Filters to appropriate fields.

f_data <- filter(data, Question %in% qfilt & YearEnd %in% yfilt & Total %in%
```

```
tfilt)

#Subset specific variables from data

sub_data <-
data.frame(f_data$YearEnd,f_data$LocationDesc,f_data$LocationAbbr,f_data$Ques
tion,f_data$Data_Value)
f_data<-drop.levels(sub_data)

#Use Spread to arrange variables by State and remove NA (Virgin Islands had
not data)

finaldata <- na.omit(spread(f_data, key = f_data.Question, value =
f_data.Data_Value))

#Clean up column names
colnames(finaldata) <-
c("YearEnd","LocDesc","LocAbbr","Overweight","Obese","NoActivity","Fruit1dy",
"Veg1dy")

finaldata<- read.csv("finaldata.csv")

# Final Clean up - Remove Column & Outliers
finaldata<-finaldata[,-1]
finaldata<-filter(finaldata, LocAbbr != "PR")
finaldata<-filter(finaldata, LocAbbr != "US")
```

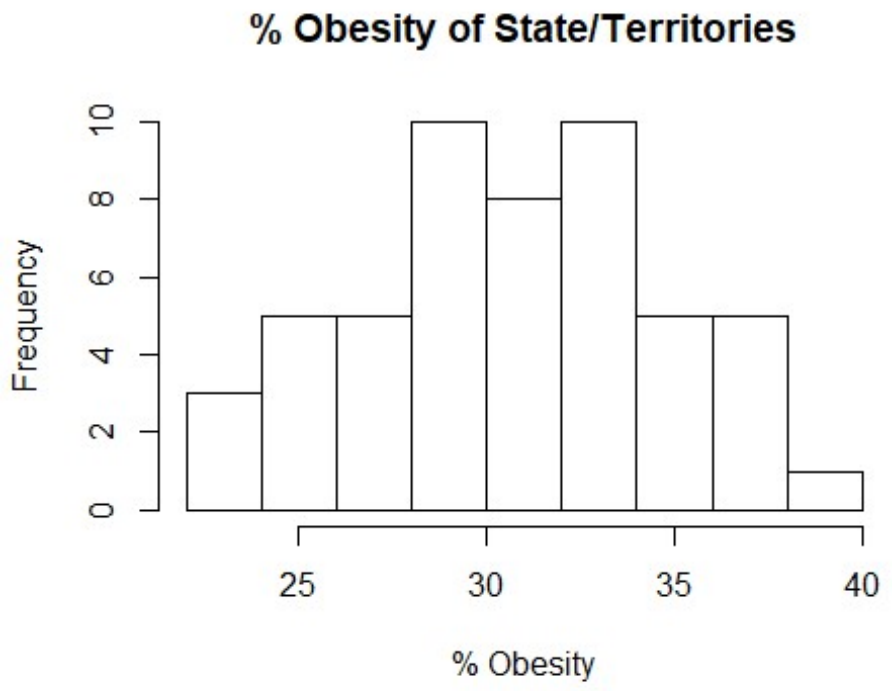
Summarize how you addressed this problem statement (the data used and the methodology employed).

The cleaned data provided 52 data points for each of the above identified variables presented at percentages of survey respondents. Additionally data was maintained related to the year and state, however was not used for further analysis. The normality of the distributions, plot the variable densities, calculate the correlation between the variables were reviewed before using linear models to determine the relationships between dependent and independent variables.

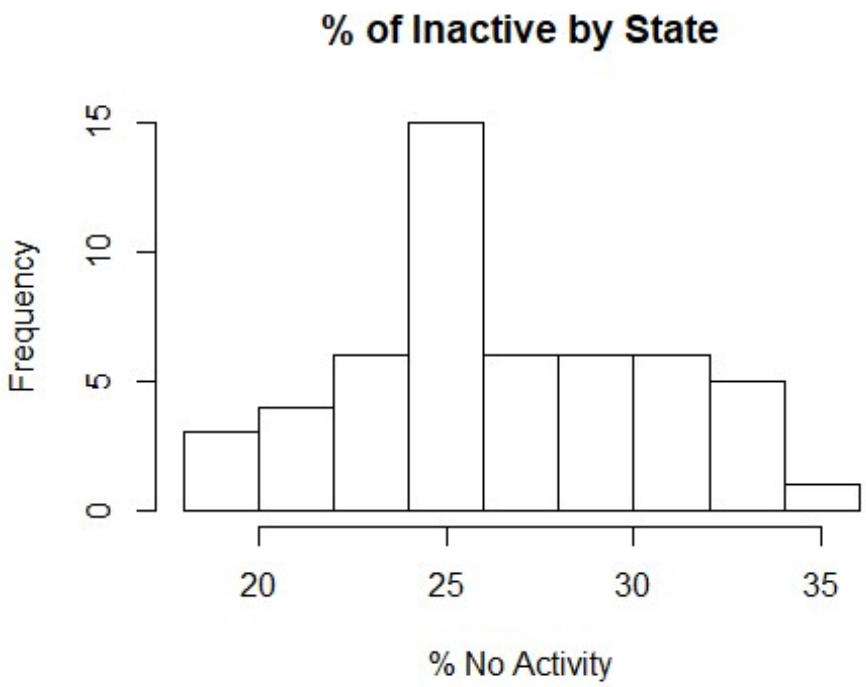
```
str(finaldata )

## 'data.frame':    52 obs. of  8 variables:
##  $ YearEnd   : int  2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 ...
##  $ LocDesc   : Factor w/ 54 levels "Alabama","Alaska",..: 1 2 3 4 5 6 7 8
9 10 ...
##  $ LocAbbr   : Factor w/ 54 levels "AK","AL","AR",..: 2 1 4 3 5 6 7 9 8 10
...
##  $ Overweight: num  33.9 32.6 35.3 35.4 35.8 36.1 36.4 36.7 31 35.6 ...
##  $ Obese     : num  36.3 34.2 29.5 35 25.1 22.6 26.9 31.8 23 28.4 ...
##  $ NoActivity: num  32 20.6 25.1 32.5 20 19.5 24 31 23 29.2 ...
##  $ Fruit1dy  : num  44.9 36.9 37.2 44.7 32.5 33 31.5 35.4 30.6 34.4 ...
##  $ Veg1dy    : num  19.3 19 20.8 19.3 21.4 17.4 16.9 17.2 13.7 19.4 ...
```

The following histograms provide a visualize depiction of the frequency of variable data.

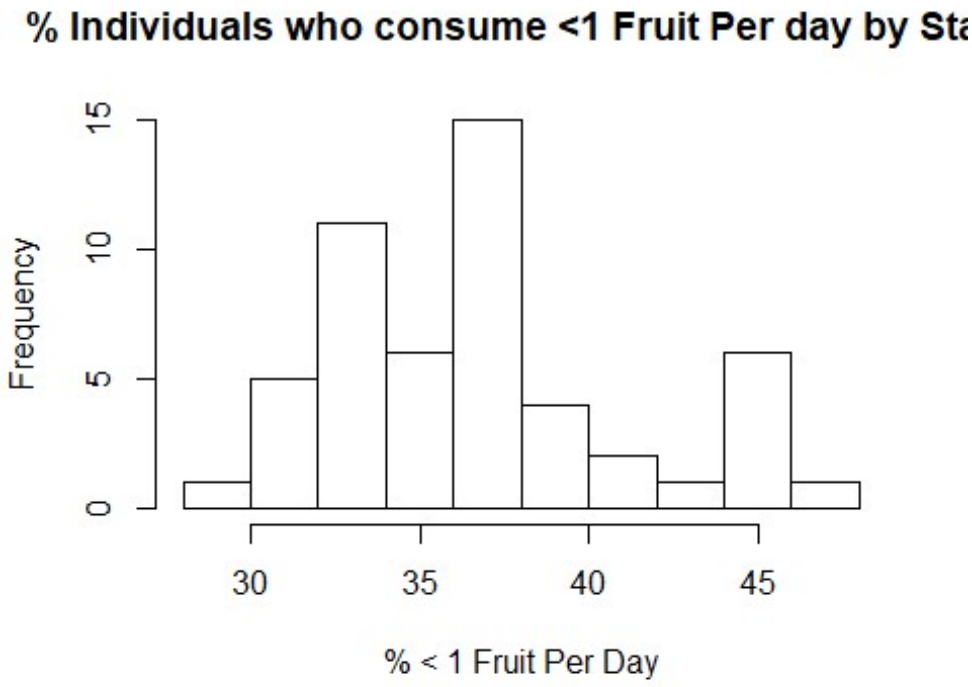
```
hist(finaldata$Obese,main="% Obesity of State/Territories",xlab="% Obesity")
```



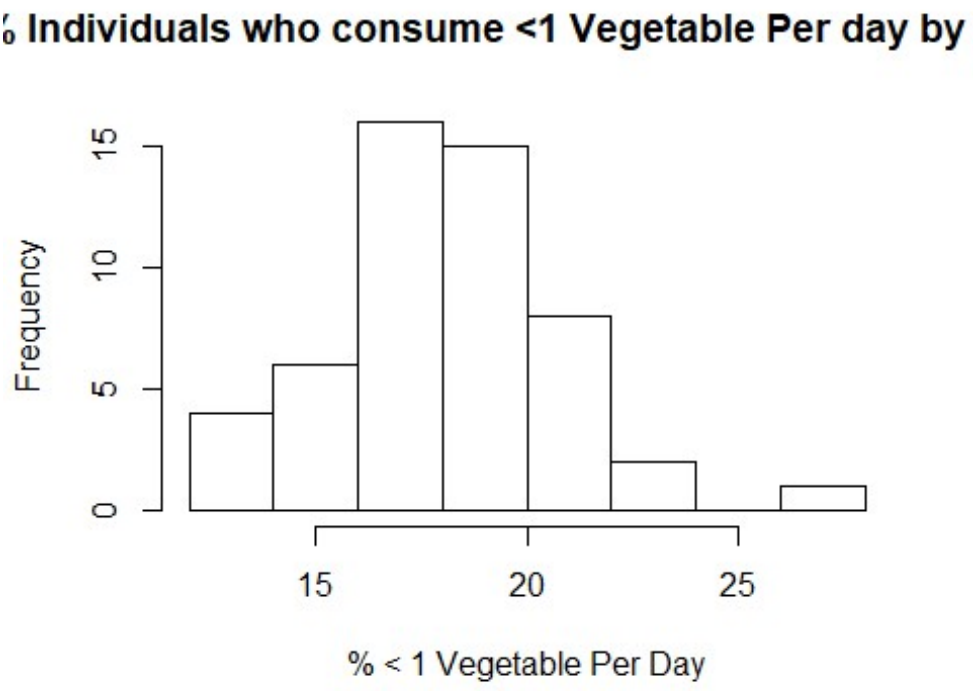

```
hist(finaldata$NoActivity,main="% of Inactive by State",xlab="% No Activity")
```




```r
hist(finaldata$Fruit1dy,main="% Individuals who consume <1 Fruit Per day by State",xlab="% < 1 Fruit Per Day")
```

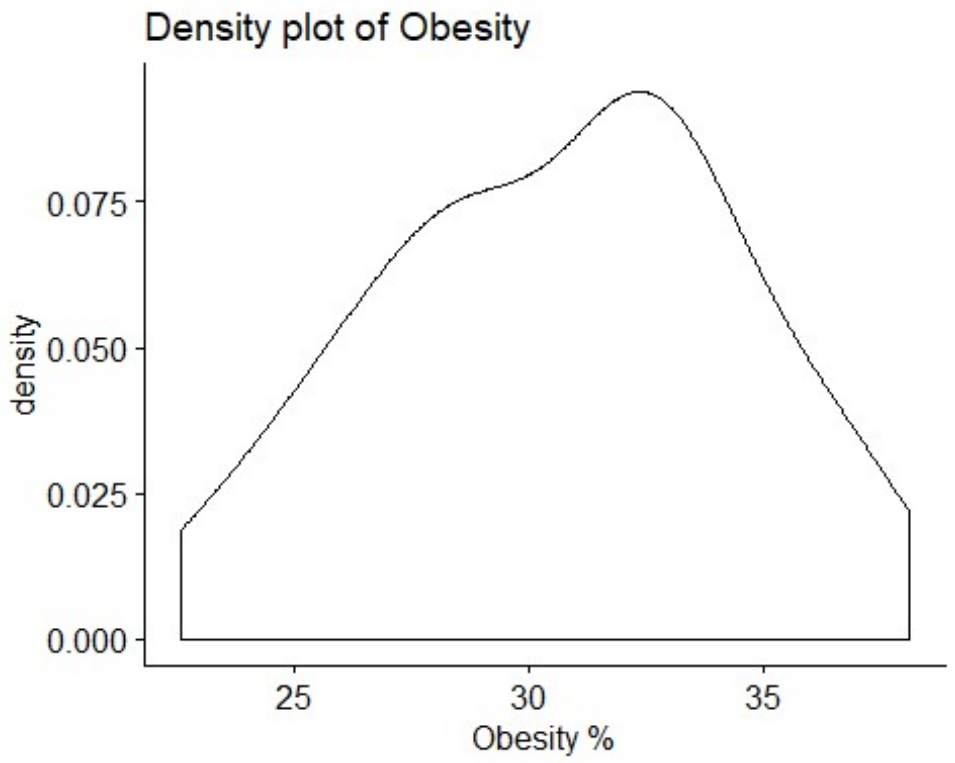


```
hist(finaldata$Veg1dy,main="% Individuals who consume <1 Vegetable Per day by
State",xlab="% < 1 Vegetable Per Day")
```

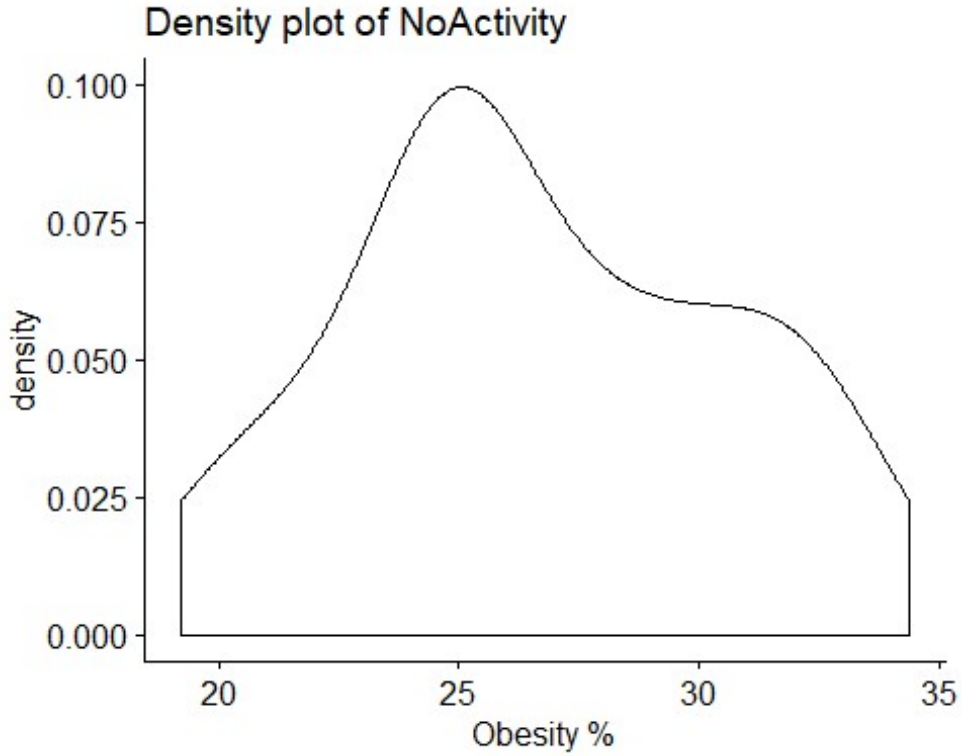


Density of the variable data is displayed below.

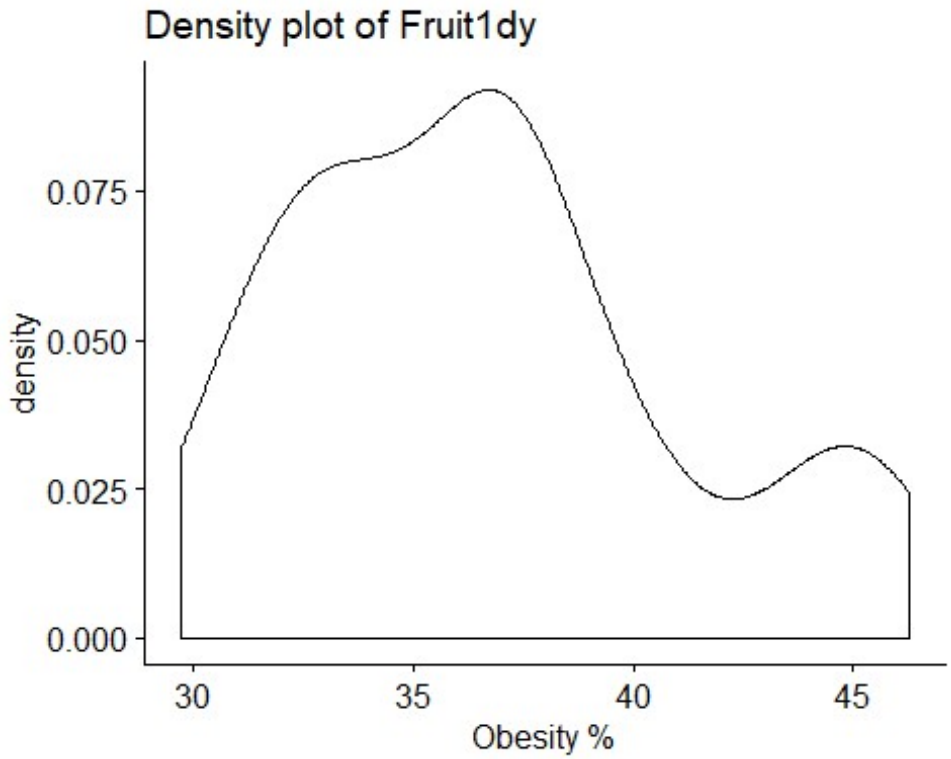
```
ggdensity(finaldata$Obese,
          main = "Density plot of Obesity",
          xlab = "Obesity %")
```



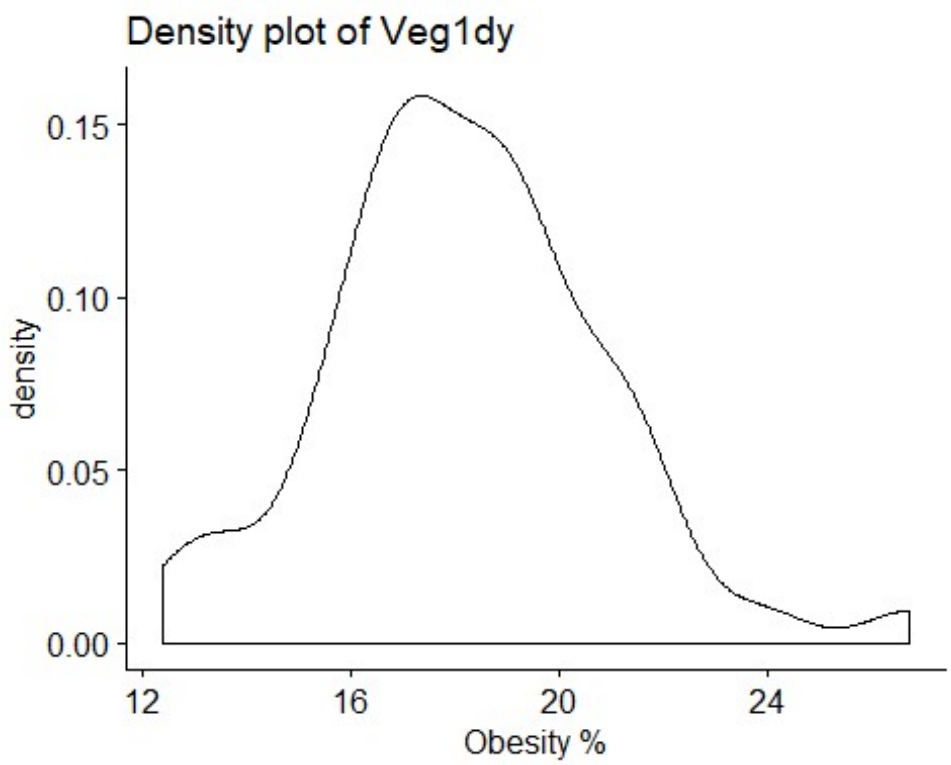

```
ggdensity(finaldata$NoActivity,
          main = "Density plot of NoActivity",
          xlab = "Obesity %")
```



```r
ggdensity(finaldata$Fruit1dy,
          main = "Density plot of Fruit1dy",
          xlab = "Obesity %")
```




```r
ggdensity(finaldata$Veg1dy,
          main = "Density plot of Veg1dy",
          xlab = "Obesity %")
```



Correlation between variables - For p-Values less than 0.05,reject the null hypothesis that the true correlation is zero (i.e. they are independent). Based on the results, reject the null hypothesis and conclude that Obesity is dependent on No Activity, Low Fruit consumption, and Low Vegetable consumption.

```
cor(finaldata$NoActivity,finaldata$Obese)
```

```
## [1] 0.669412
```

```
cor(finaldata$Fruit1dy,finaldata$Obese)
```

```
## [1] 0.7574648
```

```
cor(finaldata$Veg1dy,finaldata$Obese)
```

```
## [1] 0.2866637
```

Normality of the distributions - To determine normal distribution a p-value > 0.05 implies that the distribution of the data are not significantly different from normal distribution, or we can assume the normality. All p-values aside from Fruit are greater than 0.05, which means it is not normally distributed.

```
shapiro.test(finaldata$Obese)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  finaldata$Obese
## W = 0.98066, p-value = 0.5542
```

```
shapiro.test(finaldata$NoActivity)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  finaldata$NoActivity
## W = 0.96877, p-value = 0.1872
```

```
shapiro.test(finaldata$Fruit1dy)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  finaldata$Fruit1dy
## W = 0.9366, p-value = 0.00817
```

```
shapiro.test(finaldata$Veg1dy)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  finaldata$Veg1dy
## W = 0.97817, p-value = 0.4514
```

Correlation coefficient - Positive (closer to 1) means closer relationship, 0 means none, and negative (Closer to -1) indicates inverse relationship. Calculating the covariance in a matrix format allows interpretation of the relationship between multiple variables, in one summary output. The output indicates there is a positive relationship between Obese and % No Activity and % Less than 1 Fruit per day. A lower but still positive relationship exists between % Obese and % Less than 1 Vegetable per day. This is the weakest relationship between all variables.

```
cor(finaldata[,5:8])
```

```
##                Obese NoActivity  Fruit1dy    Veg1dy
## Obese      1.0000000  0.6694120 0.7574648 0.2866637
## NoActivity 0.6694120  1.0000000 0.8007597 0.3487646
## Fruit1dy   0.7574648  0.8007597 1.0000000 0.4604033
## Veg1dy     0.2866637  0.3487646 0.4604033 1.0000000
```

Skewness and Kurtosis of the variables - Obesity, No Activity and Vegetable are approximately symmetric while Fruit is moderately skewed. All but obesity are skewed right (positive). For kurtosis, all are generally centrally shaped with distinct tails that are not thin, but have broad, but dulled central peaks.

```
skew(finaldata[,5:8])
```

```
## [1] -0.1583748  0.1004319  0.5886527  0.3999497
```

```
kurtosi(finaldata[,5:8])
```

```
##      Obese NoActivity   Fruit1dy     Veg1dy
## -0.8039545 -0.9325643 -0.5315163  0.7844620
```

Summarize the interesting insights that your analysis provided.

Linear models were created for the obesity variable and compared to see if a model with multiple independent variables improved the fit. The Anova test indicates higher significance with the final model including more

variables, impacting the obesity outcome. Additionally, the final model indicates higher significance (p value) on the variable, Percentage of Adults consuming fruit less than one time daily on Obesity. Veg had the next highest P value and finally the Activity Component. I'd summarize the data by stating the addage you can't outrun a poor diet, as in intake, or in this case poor intake of less than 1 fruit or vegetable a day impacts obesity more than not performing physical activity.

```
test1 <- lm(Obese ~ NoActivity, data=finaldata)
test2 <- lm(Obese ~ Fruit1dy, data=finaldata)
test3 <- lm(Obese ~ Veg1dy, data=finaldata)
test4 <- lm(Obese ~ NoActivity + Fruit1dy + Veg1dy, data=finaldata)

anova(test1,test2,test3,test4)
```

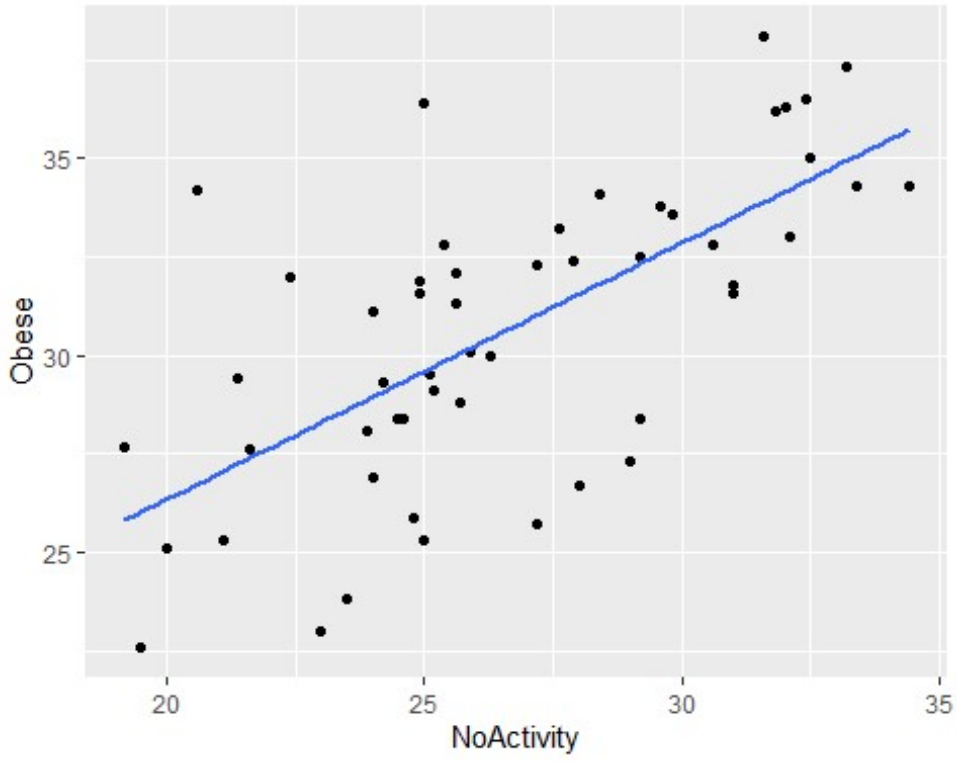
```
## Analysis of Variance Table
##
## Model 1: Obese ~ NoActivity
## Model 2: Obese ~ Fruit1dy
## Model 3: Obese ~ Veg1dy
## Model 4: Obese ~ NoActivity + Fruit1dy + Veg1dy
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     50 416.86
## 2     50 321.96  0     94.90
## 3     50 693.26  0   -371.30
## 4     48 310.34  2    382.92 29.613 4.195e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(test4)
```

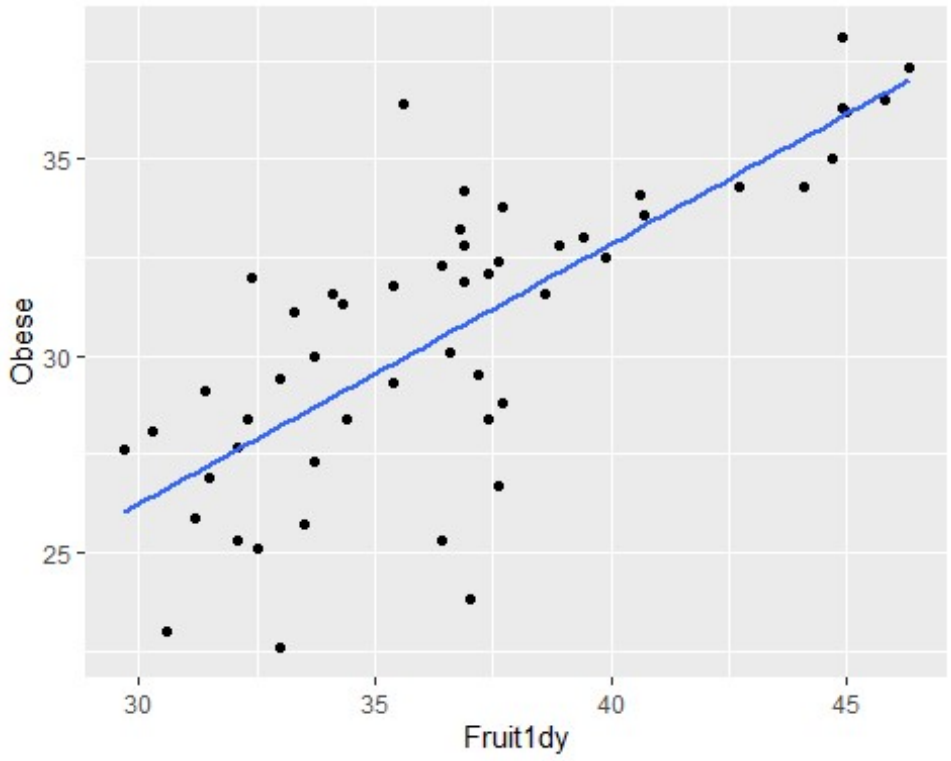
```
##
## Call:
## lm(formula = Obese ~ NoActivity + Fruit1dy + Veg1dy, data = finaldata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2187 -1.2220  0.2535  1.2148  6.7487
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.2442     3.2073   2.259  0.02849 *
## NoActivity    0.1665     0.1504   1.107  0.27387
## Fruit1dy      0.5703     0.1421   4.012  0.00021 ***
## Veg1dy       -0.1067     0.1494  -0.714  0.47878
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.543 on 48 degrees of freedom
## Multiple R-squared:  0.5891, Adjusted R-squared:  0.5635
## F-statistic: 22.94 on 3 and 48 DF,  p-value: 2.34e-09
```

The following plots show varying degrees of positive relationship between all dependent/independent variable scenarios.

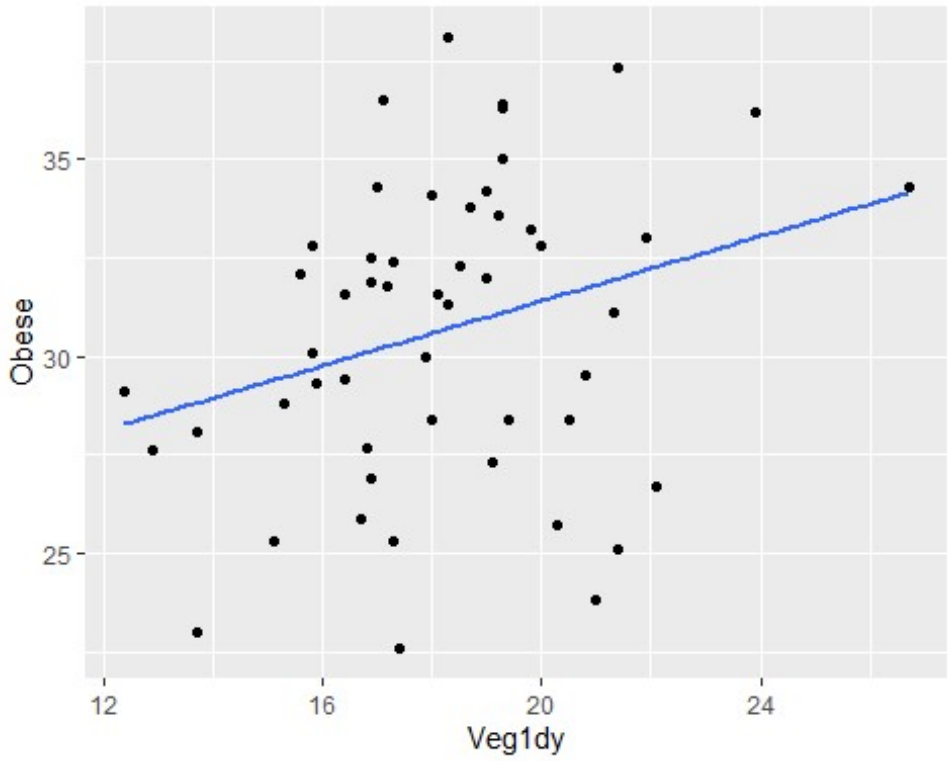
```r
# Plot linear models for specific data
ggplot(data=finaldata, aes(y=Obese, x=NoActivity))+
  geom_point() +
  geom_smooth(method="lm",se=FALSE)
```




```r
ggplot(data=finaldata, aes(y=Obese, x=Fruit1dy))+
  geom_point() +
  geom_smooth(method="lm",se=FALSE)
```




```
ggplot(data=finaldata, aes(y=Obese, x=Veg1dy))+
  geom_point() +
  geom_smooth(method="lm",se=FALSE)
```



Summarize the implications to the consumer (target audience) of your analysis.

The target audience of this analysis would be anyone concerned about limiting or preventing obesity. There were clear positive relationships between the rate of obesity of respondents and the surveyed independent variable conditions (physical activity, fruit and vegetable consumption). The lack of fruit in the group surveyed as obese was the most statistically significant variable.

Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.

Two major limitations from the data exist. The first being genetic predispositions to obesity are not surveyed for. Regardless of the relationship between obesity, physical activity, and nutrition, an individual's genetic or inherent bias towards obesity should be considered. Additionally, these the nutrition data points collected were only collected in 2017, not prior or subseqent years.