Historical Phone Usage

Arnab Panja

2020-11-16

Data Analysis

Exploratory Data Analysis Let us load the data first and observe the basic few characteritics of the data set.

```
mobile <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/20
  janitor::clean_names() %>%
  janitor::remove_empty(which = "rows")
```

The data set has 6277 observations and 7 variables. The summary of the data can be obtained using the summary function. The summary of the data is as below.

```
summary(mobile)
```

```
##     entity              code                year         total_pop
##  Length:6277        Length:6277        Min.   :1990   Min.   :5.000e+01
##  Class :character   Class :character   1st Qu.:1996   1st Qu.:4.274e+05
##  Mode  :character   Mode  :character   Median :2003   Median :4.706e+06
##                                        Mean   :2003   Mean   :2.789e+07
##                                        3rd Qu.:2010   3rd Qu.:1.648e+07
##                                        Max.   :2017   Max.   :1.359e+09
##                                                       NA's   :935
##   gdp_per_cap        mobile_subs        continent
##  Min.   :   247.4   Min.   :  0.0000   Length:6277
##  1st Qu.:  2895.4   1st Qu.:  0.5655   Class :character
##  Median :  8508.3   Median : 22.6413   Mode  :character
##  Mean   : 15832.1   Mean   : 46.4613
##  3rd Qu.: 21866.1   3rd Qu.: 88.0015
##  Max.   :135318.8   Max.   :321.8030
##  NA's   :1202       NA's   :676
```

The first few observations can be glanced as well using the head function.

```
head(mobile)
```

```
## # A tibble: 6 x 7
##   entity      code   year total_pop gdp_per_cap mobile_subs continent
##   <chr>       <chr> <dbl>     <dbl>       <dbl>       <dbl> <chr>
## 1 Afghanistan AFG    1990  13032161          NA           0 Asia
## 2 Afghanistan AFG    1991  14069854          NA           0 Asia
## 3 Afghanistan AFG    1992  15472076          NA           0 Asia
## 4 Afghanistan AFG    1993  17053213          NA           0 Asia
## 5 Afghanistan AFG    1994  18553819          NA           0 Asia
## 6 Afghanistan AFG    1995  19789880          NA           0 Asia
```

The india data in particular can be glanced as well. The use of filter helps us to select and observe the data country-wise.

```
mobile %>% filter(str_to_upper(entity) == "INDIA") %>%
  head()
```

```
## # A tibble: 6 x 7
##   entity code   year total_pop gdp_per_cap mobile_subs continent
##   <chr>  <chr> <dbl>     <dbl>       <dbl>       <dbl> <chr>
## 1 India  IND    1990 873785449       1755.     0       Asia
## 2 India  IND    1991 891910180       1738.     0       Asia
## 3 India  IND    1992 910064576       1797.     0       Asia
## 4 India  IND    1993 928226051       1845.     0       Asia
## 5 India  IND    1994 946373316       1930.     0       Asia
## 6 India  IND    1995 964486155       2037.     0.00798 Asia
```

The India data shows that the mobile subscribers were NIL till the year 1995. From 1995 there has been a growth of the mobile subsriber base as we all Indians can very well agree to this fact.

Let us see how many distinct countries by continents are present in this data set

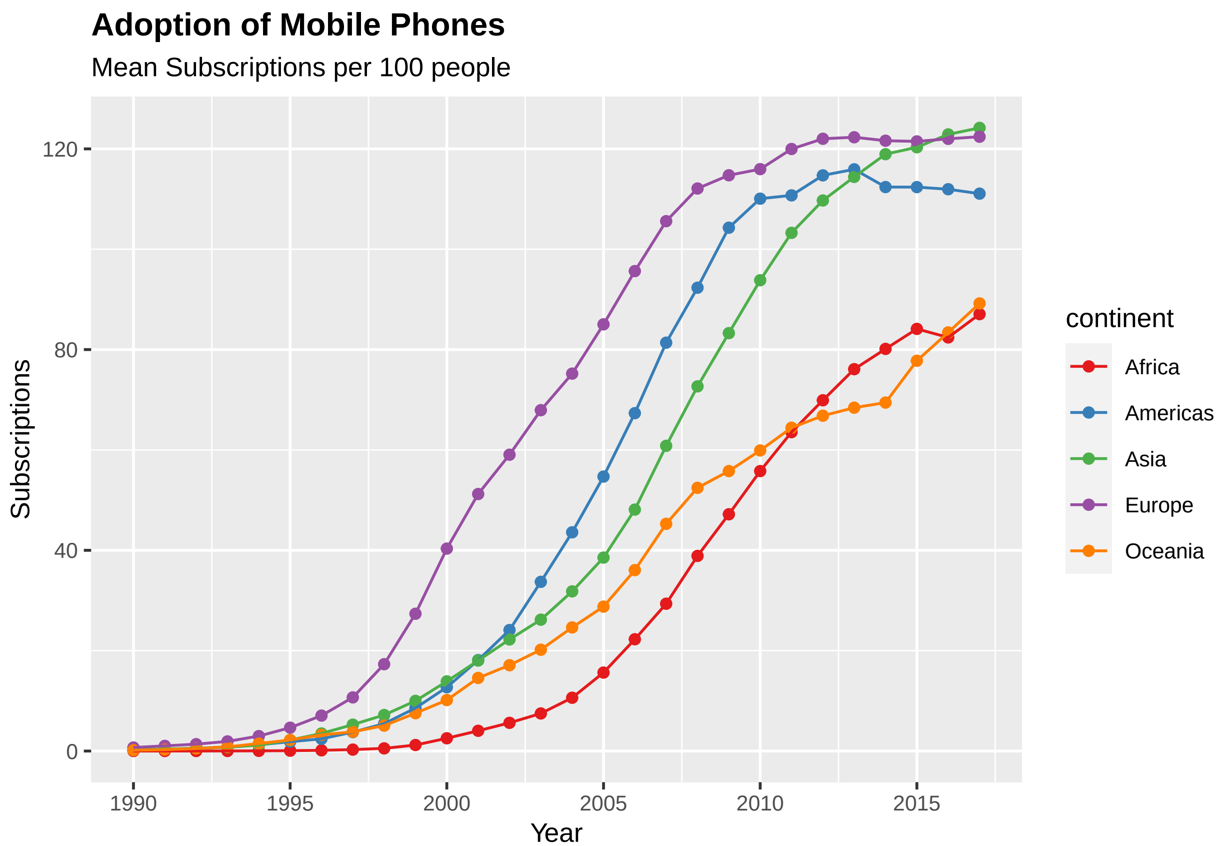
```
mobile %>% distinct(entity, continent) %>%
  group_by(continent) %>%
  summarise(country_count = n(), .groups = "drop_last") %>%
  ungroup() %>%
  arrange(-country_count)
```

```
## # A tibble: 5 x 2
##   continent country_count
##   <chr>             <int>
## 1 Africa               59
## 2 Americas             56
## 3 Asia                 54
## 4 Europe               54
## 5 Oceania              25
```

Adoption of Mobile Phones Now lets us plot the number of subscribers as a function of year for each of the countries. We will take the mean per year for every continent and then plot the mean subscribers with the year. This will give us a visualization to compare the growth of subscribers across the continents

```
mobile %>% select(year,
                  mobile_subs,
                  continent) %>%
  group_by(continent, year) %>%
  summarise(mean_subs = round(mean(mobile_subs, na.rm = TRUE), digits = 4),
            .groups = "drop_last") %>%
  ungroup() %>%
  ggplot() +
  geom_point(mapping = aes(x = year,
                           y = mean_subs,
                           color = continent),
             show.legend = TRUE) +
  geom_line(mapping = aes(x = year,
                          y = mean_subs,
                          color = continent),
             show.legend = TRUE) +
  scale_x_continuous(breaks = seq(1990, 2025, 5),
                     labels = seq(1990, 2025, 5)) +
```

```
  scale_color_brewer(palette = "Set1") +
  theme(text = element_text(size = 10),
        plot.title = element_text(face = "bold")) +
  labs(x = "Year",
       y = "Subscriptions",
       title = "Adoption of Mobile Phones",
       subtitle = "Mean Subscriptions per 100 people")
```



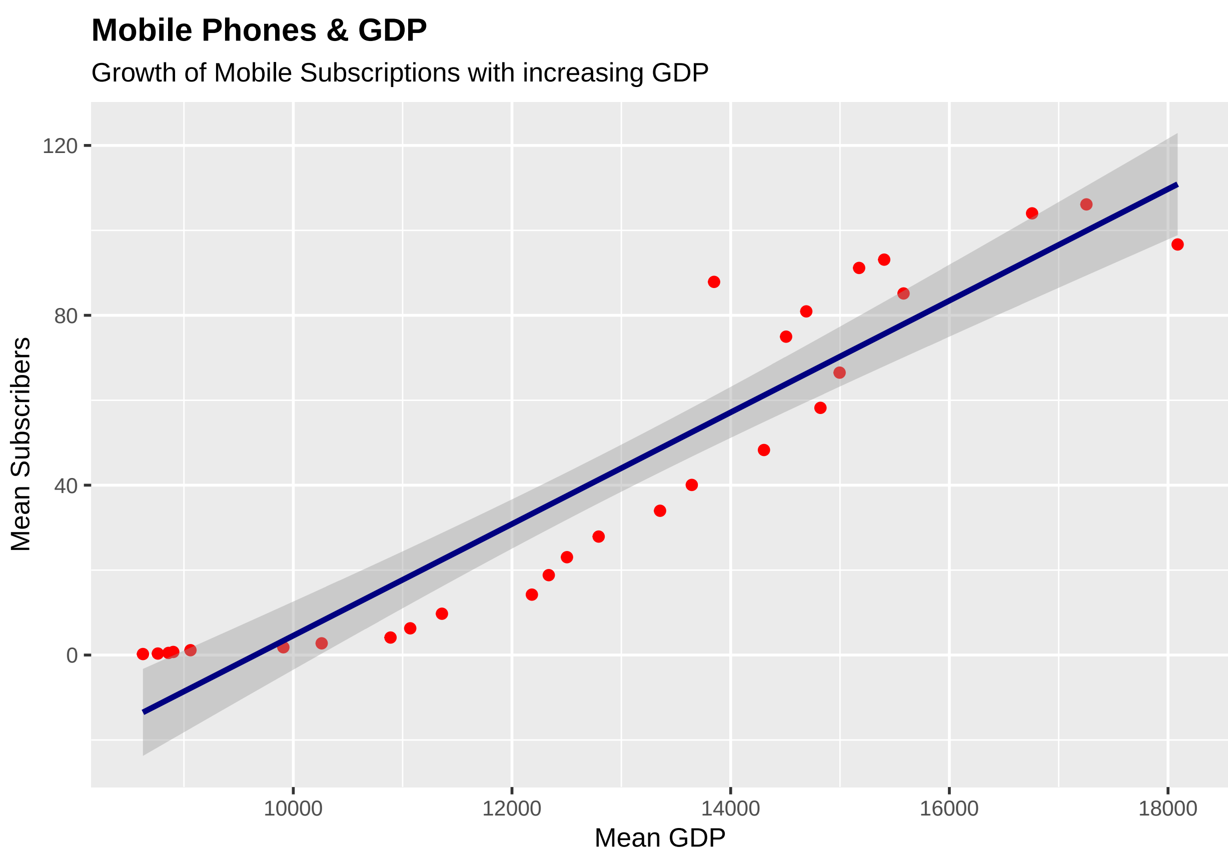
The graphic above shows how mobile phones have been adopted in the continents. Growth of the adoption of mobile phones have been in the following sequence:-

1. Europe
2. Americas
3. Asia
4. Oceania
5. Africa

Subscribers & Mean GDP per capita at PPP The number of subscribers can be a function of GDP and the Population of a particular period of time. Lets see the variation of mobile subscribers with the mean GDP of the world as well with mean population of the world.

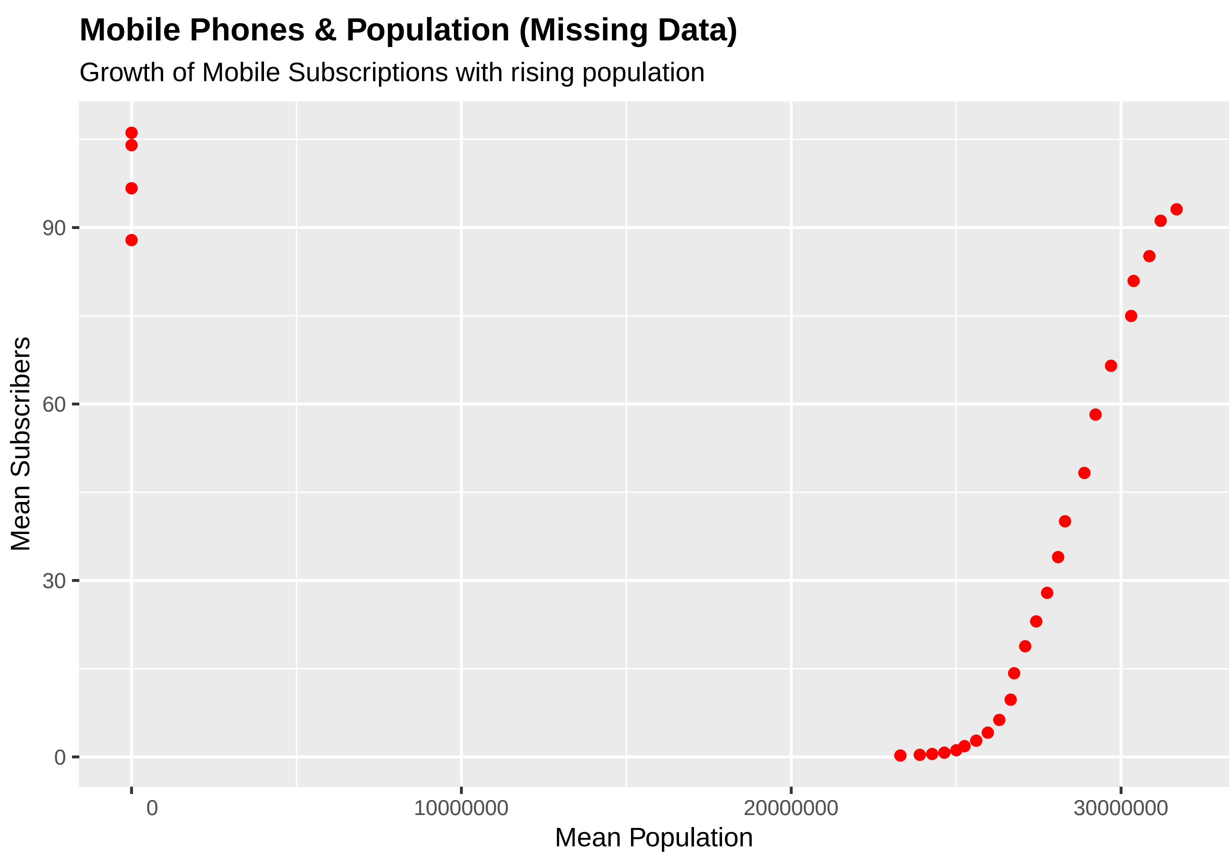
```
mobile %>% select(year,
                  mobile_subs,
                  gdp_per_cap) %>%
  mutate(gdp_per_cap = replace_na(gdp_per_cap, 0),
         mobile_subs = replace_na(mobile_subs, 0)) %>%
  group_by(year) %>%
```

```r
summarise(mean_gdp = mean(gdp_per_cap),
          mean_subs = mean(mobile_subs, na.rm = TRUE),
          .groups = "drop_last") %>%
ungroup() %>%
arrange(year) %>%
ggplot() +
geom_point(mapping = aes(x = mean_gdp,
                         y = mean_subs),
          show.legend = FALSE,
          color = "red") +
geom_smooth(mapping = aes(x = mean_gdp,
                          y = mean_subs),
            method = "lm",
            formula = "y ~ x",
            color = "navyblue") +
theme(text = element_text(size = 10),
      plot.title = element_text(face = "bold")) +
labs(x = "Mean GDP",
     y = "Mean Subscribers",
     title = "Mobile Phones & GDP",
     subtitle = "Growth of Mobile Subscriptions with increasing GDP")
```



Subscribers & Mean Population Now let us see the growth of mobile subscriptions with the growing population of the world. With a growing population there will normally be a greater demand of mobile phones. The reason being there will be a greater need of communication with a growing population. GDP also plays a part in this. Let us study the variation below using some plots.

```
mobile %>% select(year,
                  mobile_subs,
                  total_pop) %>%
  mutate(total_pop = replace_na(total_pop, 0),
         mobile_subs = replace_na(mobile_subs, 0)) %>%
  group_by(year) %>%
  summarise(mean_pop = mean(total_pop),
            mean_subs = mean(mobile_subs, na.rm = TRUE),
            .groups = "drop_last") %>%
  ungroup() %>%
  arrange(year) %>%
  ggplot() +
  geom_point(mapping = aes(x = mean_pop,
                           y = mean_subs),
             show.legend = FALSE,
             color = "red") +
  scale_x_continuous(labels = function(x) format(x, scientific = FALSE)) +
  theme(text = element_text(size = 10),
        plot.title = element_text(face = "bold")) +
  labs(x = "Mean Population",
       y = "Mean Subscribers",
       title = "Mobile Phones & Population (Missing Data)",
       subtitle = "Growth of Mobile Subscriptions with rising population")
```



The graphic above shows that the mean subscribers are recorded for a set of observations with mean population as zero (the points on the top left of the above graphic). This looks like a data quality issue where the population data have not been recorded and replacing them with zero has not been correct. Let us see which of the records have this issue and examine if there is a better way to fill the missing population values.

```
mobile %>% select(year,
                  mobile_subs,
                  total_pop) %>%
  mutate(total_pop = replace_na(total_pop, 0),
         mobile_subs = replace_na(mobile_subs, 0)) %>%
  group_by(year) %>%
  summarise(mean_pop = mean(total_pop),
            mean_subs = mean(mobile_subs, na.rm = TRUE),
            .groups = "drop_last") %>%
  ungroup() %>%
  filter(mean_pop == 0, mean_subs != 0)
```

```
## # A tibble: 4 x 3
##    year mean_pop mean_subs
##   <dbl>    <dbl>     <dbl>
## 1  2014        0     104.
## 2  2015        0      87.9
## 3  2016        0     106.
## 4  2017        0      96.7
```

Let us inspect a bit further to see which observations have resulted in this issue.

```
mobile %>% select(year,
                  mobile_subs,
                  total_pop) %>%
  filter(year == 2014) %>%
  head()
```

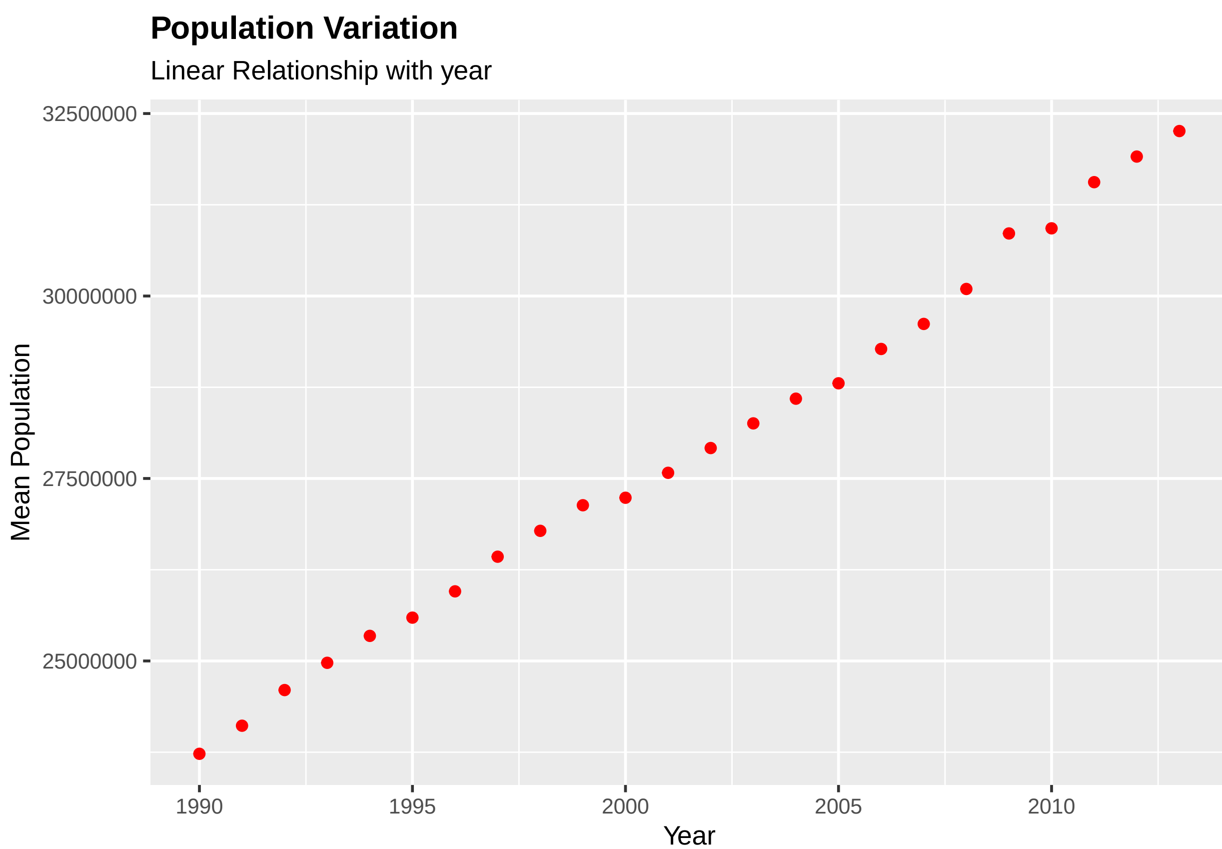
```
## # A tibble: 6 x 3
##    year mobile_subs total_pop
##   <dbl>       <dbl>     <dbl>
## 1  2014        56.2        NA
## 2  2014       115.         NA
## 3  2014       111.         NA
## 4  2014        83.6        NA
## 5  2014        52.2        NA
## 6  2014       121.         NA
```

So we see that population data has not been recorded for the years 2014, 2015, 2016, 2017 as we predicted.

So let us now quickly see how the mean population has been varying with year till 2013.

```
pop_df <- mobile %>% select(year, total_pop) %>%
  filter(year <= 2013) %>%
  group_by(year) %>%
  summarise(mean_pop = mean(total_pop, na.rm = TRUE),
            .groups = "drop_last") %>%
  arrange(year)


ggplot(data = pop_df, mapping = aes(x = year, y = mean_pop)) +
  geom_point(color = "red", show.legend = FALSE) +
  theme(text = element_text(size = 10),
```

```
        plot.title = element_text(face = "bold")) +
  labs(x = "Year",
       y = "Mean Population",
       title = "Population Variation",
       subtitle = "Linear Relationship with year")
```



Statistical Modelling - Missing Population Data

Linear Regression Since the relationship appears to be a linear one we can fit a linear model and predict the mean populations for the missing years 2014, 2015, 2016 and 2017. So let us do that now.

```
lm_model <- lm(formula = mean_pop ~ year,
               data = pop_df)

summary(lm_model)
```

```
##
## Call:
## lm(formula = mean_pop ~ year, data = pop_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -349141 -145094    6172  138578  268233
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -690410536   10947671  -63.06   <2e-16 ***
## year            358885       5470   65.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 185500 on 22 degrees of freedom
## Multiple R-squared:  0.9949, Adjusted R-squared:  0.9947
## F-statistic:  4305 on 1 and 22 DF,  p-value: < 2.2e-16
```

The summary of the linear model gives a high value of R-Squared meaning the model is quite reliable for predicting mean population for the missing years. A high value of the R-Squared/Adjusted R-Squared or a very low p-value are indicative of the model being a good fit of the data.

We now use this model to predict the mean population of the missing years 2014, 2015, 2016 and 2017.

```
# create test data frame
test_df <- data.frame(year = c(2014:2017),
                        stringsAsFactors = FALSE)

# predict the missing population values
predict_pop <- predict(object = lm_model, newdata = test_df)

new_pop_df <- as_tibble(cbind("year" = test_df$year,
                              mean_pop = predict_pop))

new_pop_df
```

```
## # A tibble: 4 x 2
##    year  mean_pop
##   <dbl>     <dbl>
## 1  2014 32383496.
## 2  2015 32742381.
## 3  2016 33101265.
## 4  2017 33460150.
```

So now having calculated the mean population for the missing years we go back to get the plot for the variation of population with subscribers for all years till 2014 and beyond as far as the population values were recorded.

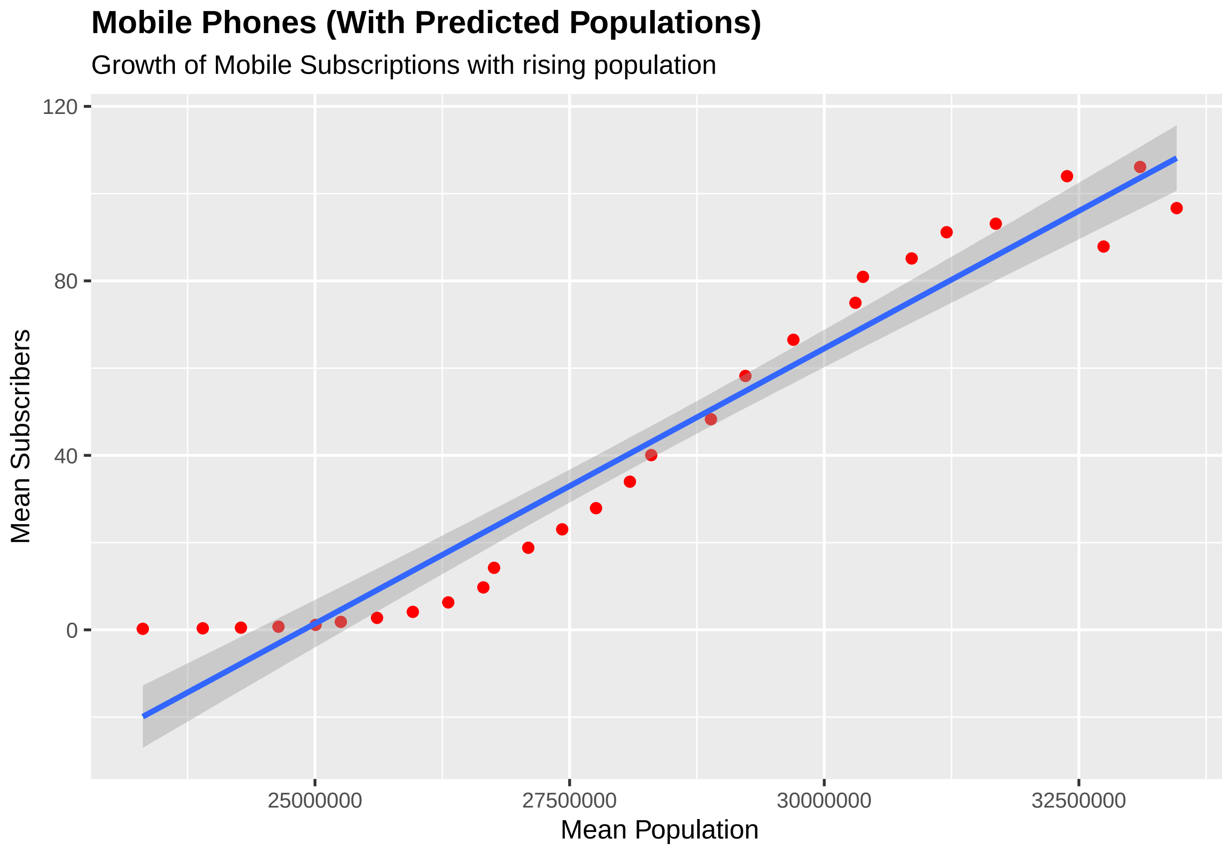
```
pop_study_1 <- mobile %>% select(year,
                  mobile_subs,
                  total_pop) %>%
  filter(year < 2014 | year > 2017) %>%
  mutate(total_pop = replace_na(total_pop, 0),
          mobile_subs = replace_na(mobile_subs, 0)) %>%
  group_by(year) %>%
  summarise(mean_pop = mean(total_pop),
            mean_subs = mean(mobile_subs, na.rm = TRUE),
            .groups = "drop_last") %>%
  ungroup() %>%
  arrange(year)

# combine the missing values as well
pop_study_2 <- bind_cols(mobile %>% select(year,
                  mobile_subs,
                  total_pop) %>%
  filter(year >= 2014 & year <= 2017) %>%
```

```r
  mutate(mobile_subs = replace_na(mobile_subs, 0)) %>%
  group_by(year) %>%
  summarise(mean_subs = mean(mobile_subs),
            .groups = "drop_last") %>%
  ungroup() %>%
  arrange(year), "mean_pop" = new_pop_df$mean_pop) %>%
  select(year, mean_pop, mean_subs)

# combined data set with all filled up population values
pop_study_comb <- bind_rows(pop_study_1, pop_study_2)

head(pop_study_comb)
```

```
## # A tibble: 6 x 3
##    year  mean_pop mean_subs
##   <dbl>     <dbl>     <dbl>
## 1  1990 23309651.     0.236
## 2  1991 23898319.     0.351
## 3  1992 24272974.     0.499
## 4  1993 24641571.     0.721
## 5  1994 25005575.     1.14
## 6  1995 25253783.     1.83
```

Now having prepared the data frame after predicting and adding the predicted values back for the missing mean populations we take a look at the plot once again as below for mean subscribers and its growth with the mean population.

```r
ggplot(data = pop_study_comb) +
  geom_point(mapping = aes(x = mean_pop,
                           y = mean_subs),
            show.legend = FALSE,
            color = "red") +
  geom_smooth(mapping = aes(x = mean_pop,
                            y = mean_subs),
              method = "lm",
              formula = "y ~ x",
              show.legend = FALSE) +
  scale_x_continuous(labels = function(x) format(x, scientific = FALSE)) +
  theme(text = element_text(size = 10),
        plot.title = element_text(face = "bold")) +
  labs(x = "Mean Population",
       y = "Mean Subscribers",
       title = "Mobile Phones (With Predicted Populations)",
       subtitle = "Growth of Mobile Subscriptions with rising population")
```



So this analysis gives a very good insight into how during an analysis we can identify some missing observations and how the nature of the variables can be studied to create a statistical model to predict the missing values. These predicted values can then be substituted for the missing values with the original full data and the analysis can be proceeded based on that with a reasonable degree of accuracy.

This notebook therefore demonstrates the following three critical aspects of data science

1. Exploratory Data Analysis
2. Data Visualization
3. Identification of missing values
4. Prediction of the missing values using a statistical model