Topic:

1. Please analyse using visual techniques #economics dataset available in package # ggplot2.
2. Visualize characteristics of this dataset.
3. Create bubble charts for 3 variables of this dataset.

A. Data Understanding

- A data frame with 574 rows and 6 variables:
    - Date: Month of Data collected from 01/07/1967 to 01/04/2015 (DD/MM/YYYY)
    - Pce: personal consumption expenditures, in billions of dollars
    - Pop: total population, in thousands
    - Psavert: personal savings rate
    - Uempmed: median duration of unemployment, in weeks
    - Unemploy: number of unemployed in thousands

- Before analyzing, we take a look at which data type within the dataset and also their overview values.

```
> str(a)
spc_tbl_ [574 × 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ date    : Date[1:574], format: "1967-07-01" "1967-08-01" ...
 $ pce     : num [1:574] 507 510 516 512 517 ...
 $ pop     : num [1:574] 198712 198911 199113 199311 199498 ...
 $ psavert : num [1:574] 12.6 12.6 11.9 12.9 12.8 11.8 11.7 12.3 11.7 12.3 ...
 $ uempmed : num [1:574] 4.5 4.7 4.6 4.9 4.7 4.8 5.1 4.5 4.1 4.6 ...
 $ unemploy: num [1:574] 2944 2945 2958 3143 3066 ...
```

```
> summary(a)
      date                pce              pop            psavert
 Min.   :1967-07-01   Min.   :  506.7   Min.   :198712   Min.   : 2.200
 1st Qu.:1979-06-08   1st Qu.: 1578.3   1st Qu.:224896   1st Qu.: 6.400
 Median :1991-05-16   Median : 3936.8   Median :253060   Median : 8.400
 Mean   :1991-05-17   Mean   : 4820.1   Mean   :257160   Mean   : 8.567
 3rd Qu.:2003-04-23   3rd Qu.: 7626.3   3rd Qu.:290291   3rd Qu.:11.100
 Max.   :2015-04-01   Max.   :12193.8   Max.   :320402   Max.   :17.300
    uempmed          unemploy
 Min.   : 4.000   Min.   : 2685
 1st Qu.: 6.000   1st Qu.: 6284
 Median : 7.500   Median : 7494
 Mean   : 8.609   Mean   : 7771
 3rd Qu.: 9.100   3rd Qu.: 8686
 Max.   :25.200   Max.   :15352
```

All 6 variables are numerical category

- Then, checking for missing and invalid data

```
> colSums(is.na(a))
    date      pce      pop  psavert  uempmed unemploy
       0        0        0        0        0        0
```

There is no missing value in this dataset.

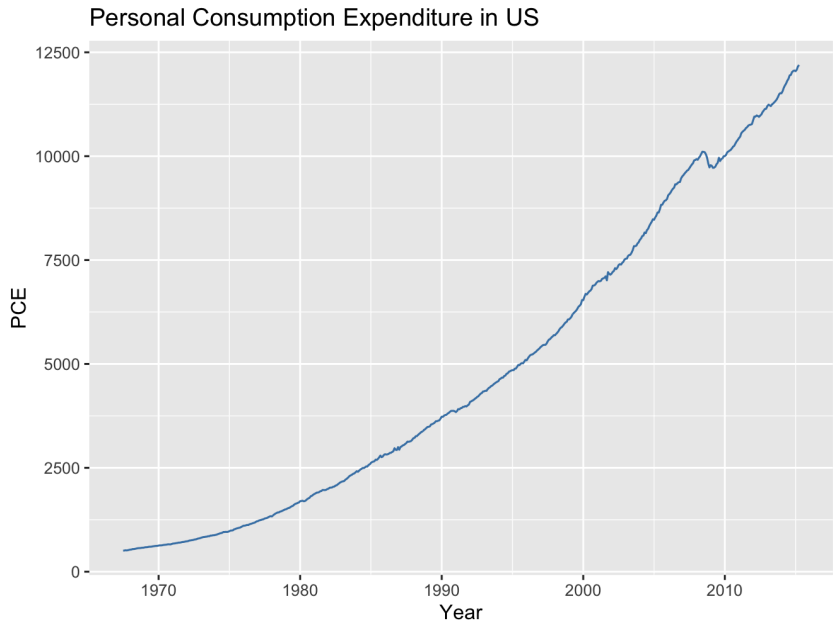
B. Methodology

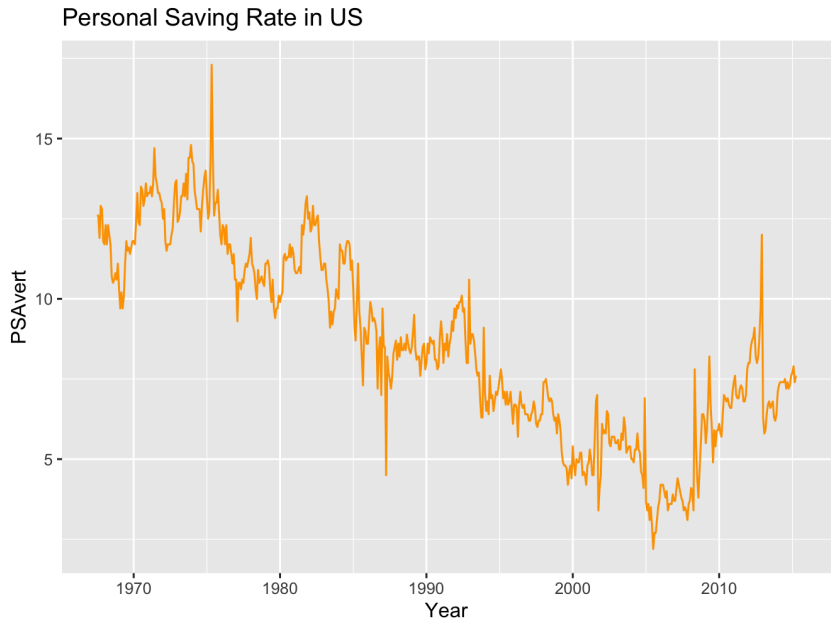
In order to better visualize characteristics for the economics dataset, we could look further into the relationship between variables.

1. Time-series analysis: using variable date, visualising trend of pce, psavert, uempmed, unemploy through years.
2. Histogram: to see distribution and outliers of variables.
3. Using Bubble Chart to analyze relationships between variables on multiple dimensions.

C. Data Analysis

1. Time-series analysis





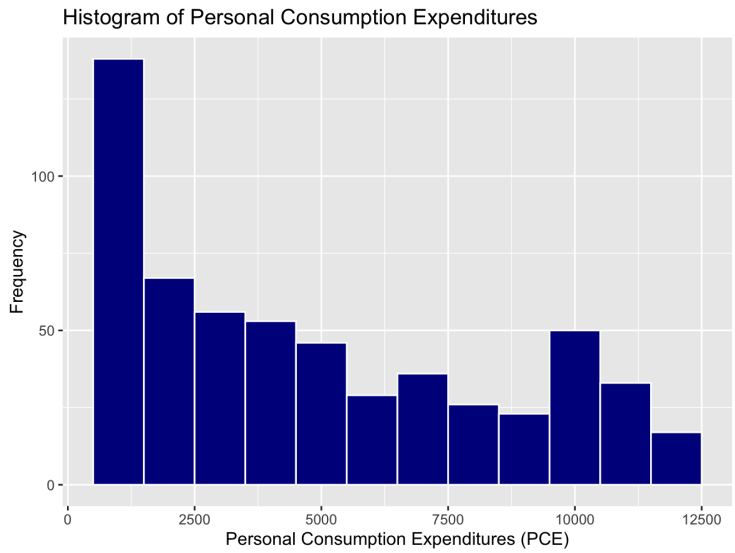
Graph 1 & 2.

Comparing PCE and PSR through the years, they have opposite patterns, while PCE was gradually increasing, PSR fluctuated downward.
It is concluded that US citizens shifted their expenses to purchasing goods instead of keep it in the bank for saving since 1967.

Graph 3 & 4.

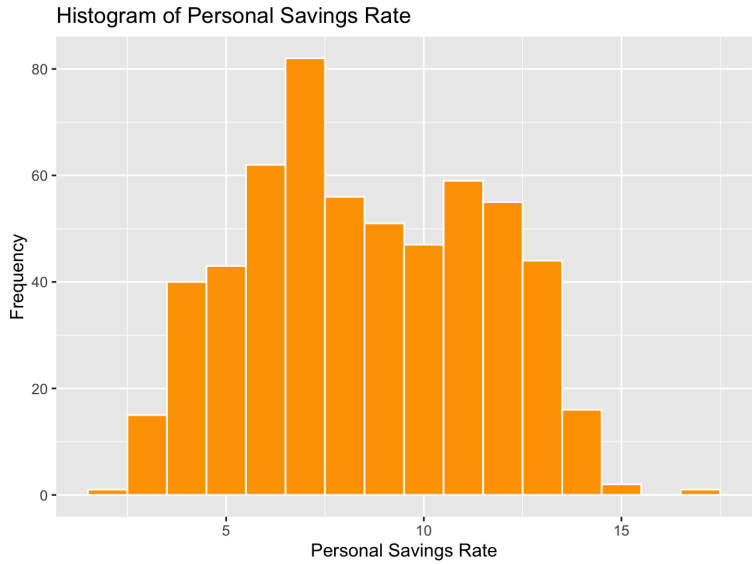
Mentioning unemployment factor in US's economics, over 15.000 unemployed were recorded in 2009, reaching its peak. In the following year, the median duration of unemployment also reached the peak at 25.2 weeks. This might be the post-recession results causing bad effect to the employment market.

2. Histogram chart



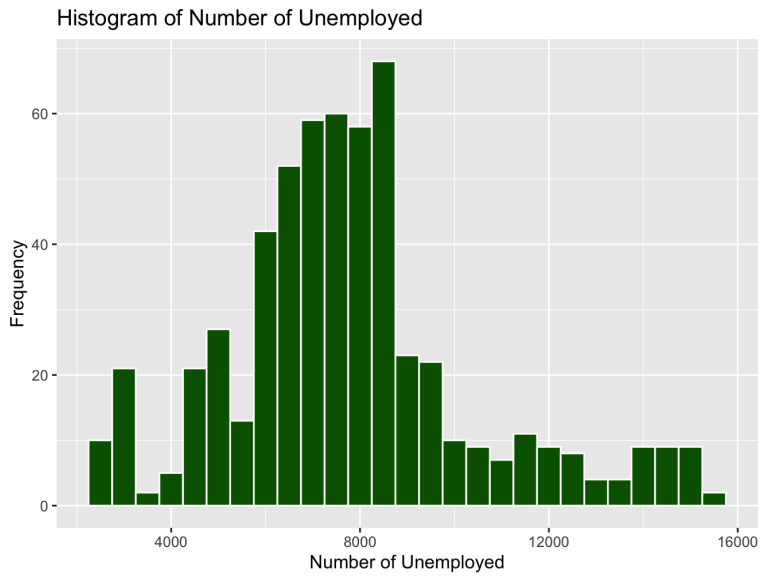
Graph 5

PCE values are most concentrated in the 0-1250 billion dollars range, indicating a left-skewed data distribution.



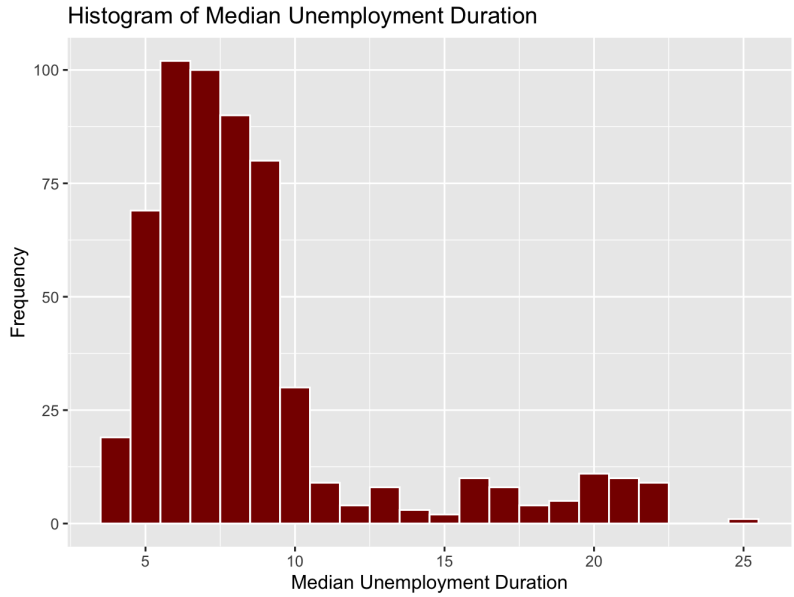
Graph 6

Psavert values are distributed fairly evenly, however its peak is to the left at around 7.5%, the number of savings above 15% is very few and almost non-existent.



Graph 7

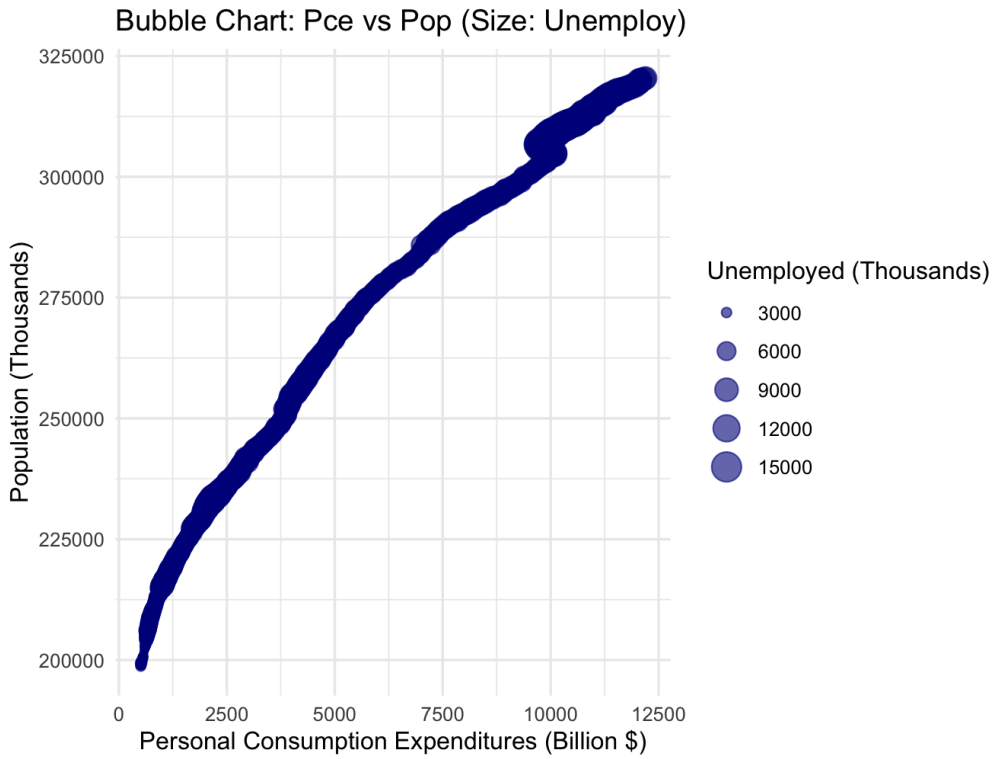
The number of more than 9,000 unemployed people was the highest recorded in the period.



Graph 8

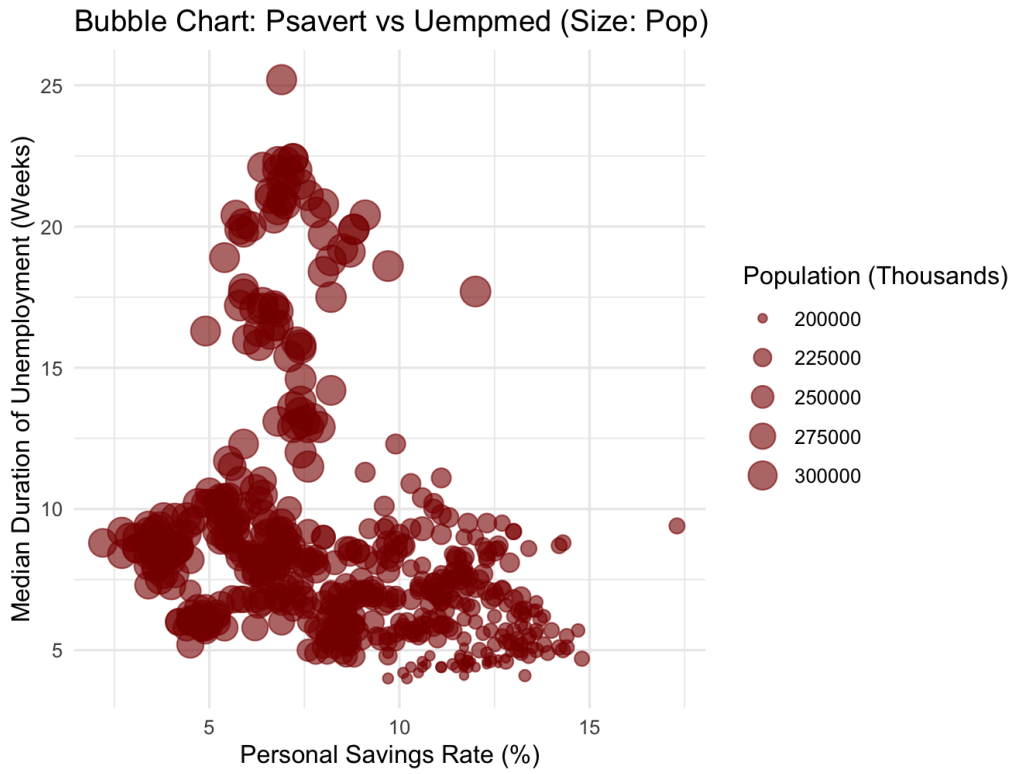
The histogram is skewed to the left, indicating that unemployment periods ranging from 5 to 10 weeks are most common.

3. Bubble chart



Graph 9

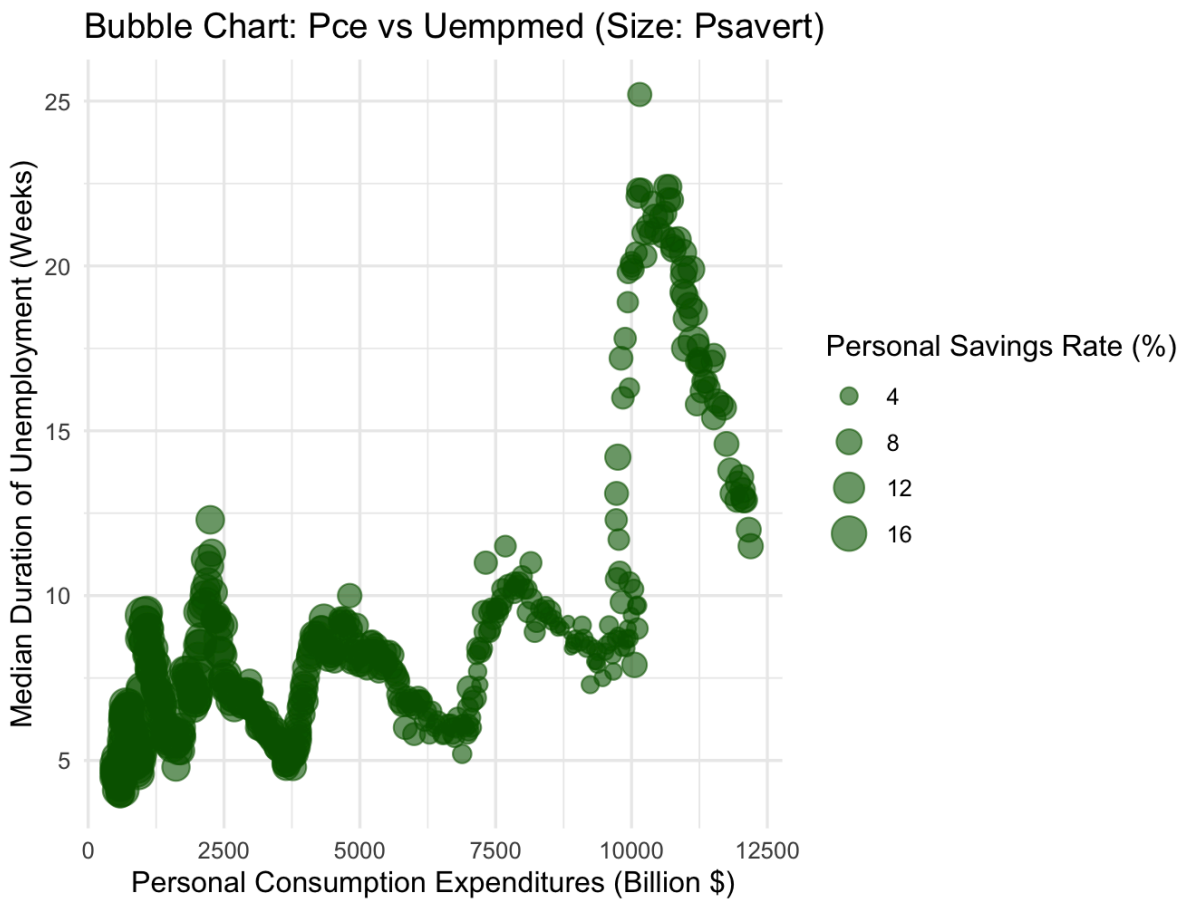
Population had a clear positive linear relationship with PCE over the observed period.
The higher the population, the higher the PCE.
With the size of the bubble not being too different, the number of unemployed people is quite even over the years, fluctuating between 9,000-12,000; however, it has an increasing pattern over time, the beginning of the period records around 3,000 unemployed people, that number increases about 4 times by the end of the period. Another point worth noting is that when the Population reaches more than 300,000 people, the PCE falls to around 10,000, the number of unemployed people tends to increase sharply (more than 15 thousand people).



Graph 10

The extremes are in the low saving rates range of 5-10% but contain the highest Uempmed indices (around 25 weeks). With the data fluctuating the number of weeks of unemployment from 5-10 being the most recorded, their Psavert ranges from about 2.5-15%.

The largest bubble sizes are recorded in the high Uempmed range of 15 weeks and above, and the low Psavert (below 5%). This shows that the larger the population, the longer the unemployment period and the lower the saving rate.



Graph 11

With low unemployment periods, from 2-12 weeks, the PCE index fluctuates around $10 billion. What is special about this chart is that with high Uempmed, (over 15 weeks), PCE increases more than with low Unempmed; PCE is recorded at levels above $10 billion. Combined with Graph 6 above, when Uempmed is high over 15 weeks, only Psavert decreases significantly, we can assume that it is due to the lag between unemployment and spending changes. Because personal spending may not decrease immediately when unemployment is prolonged. People often use savings or borrow before having to reduce spending; this leads to Pce remaining high in the early or middle stages of prolonged unemployment and Psavert decreasing.

The bubble size is largest during the low Uempmed period (5-10 weeks) and with PCE below $5 billion, which is explained by the fact that when spending is reduced and unemployment is short, people tend to save more.

D. Conclusion

When predicting PCE values, we should look at possible variables such as Psavert, Unemploy, Pop, Uempmed instead of just analyzing a single variable that can affect PCE, because each variable will return a different correlation and give misleading results if we only analyze based on a single variable.
Population can be a very important factor when we analyze a country's economy, but it also needs to be combined with other factors to make an accurate assessment.