Andrew Alford
12/10/2023
IST 686 Quantitative Reasoning for Data Science

Homework 10

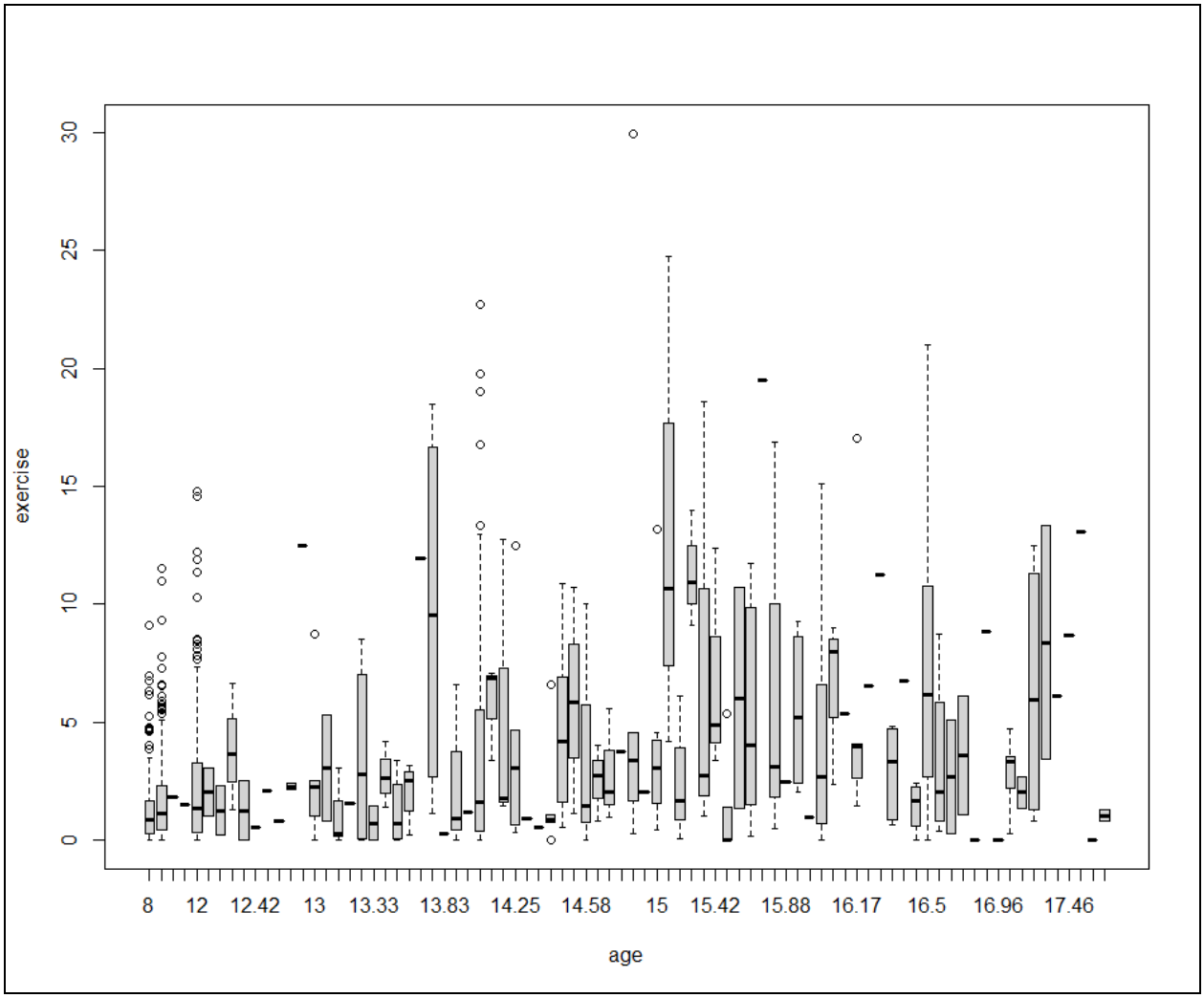
1. Download and library the nlme package and use data (“Blackmore”) to activate the Blackmore data set. Inspect the data and create a box plot showing the exercise level at different ages. Run a repeated measures ANOVA to compare exercise levels at ages 8, 10, and 12 using aov(). You can use a command like, myData <- Blackmore[Blackmore$age<=12,], to subset the data. Keeping in mind that the data will need to be balanced before you can conduct this analysis, try running a command like this, table(myData$subject,myData$age)), as the starting point for cleaning up the data set.

First things first, I downloaded the nlme package, used the library() function to initialize it, and got a view of the Blackmore data set. To do this, I used the code below:

```
#Question 1

install.packages("nlme")
library(nlme)
library(car)
data("Blackmore")
view(Blackmore)
?Blackmore
```

Next, I used a simple boxplot() function showing the exercise levels at different ages:



Being completely honest, I had a lot of trouble balancing and running the repeated measures ANOVA test. I was luckily able to get some help from my friend and colleague Bill Steele, who is also in this class. With his help, I used the following code to run the aov() function:

```
#Running the ANOVA
myData <- Blackmore[Blackmore$age %in% c(8, 10, 12),]
table(myData$subject, myData$age)
complete_subjects <- myData[ave(myData$exercise, myData$subject, FUN = length) == 3,]
complete_subjects <- droplevels(complete_subjects)
table(complete_subjects$subject, complete_subjects$age)
aov_results <- aov(exercise ~ age + Error(subject/age), data = complete_subjects)
summary(aov_results)
```

With this I got the following result:

```
Error: subject
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 175   1941   11.09

Error: subject:age
          Df Sum Sq Mean Sq F value   Pr(>F)
age        1   99.4   99.43   39.89 2.15e-09 ***
Residuals 175  436.2    2.49
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Error: Within
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 176  197.5   1.122
```

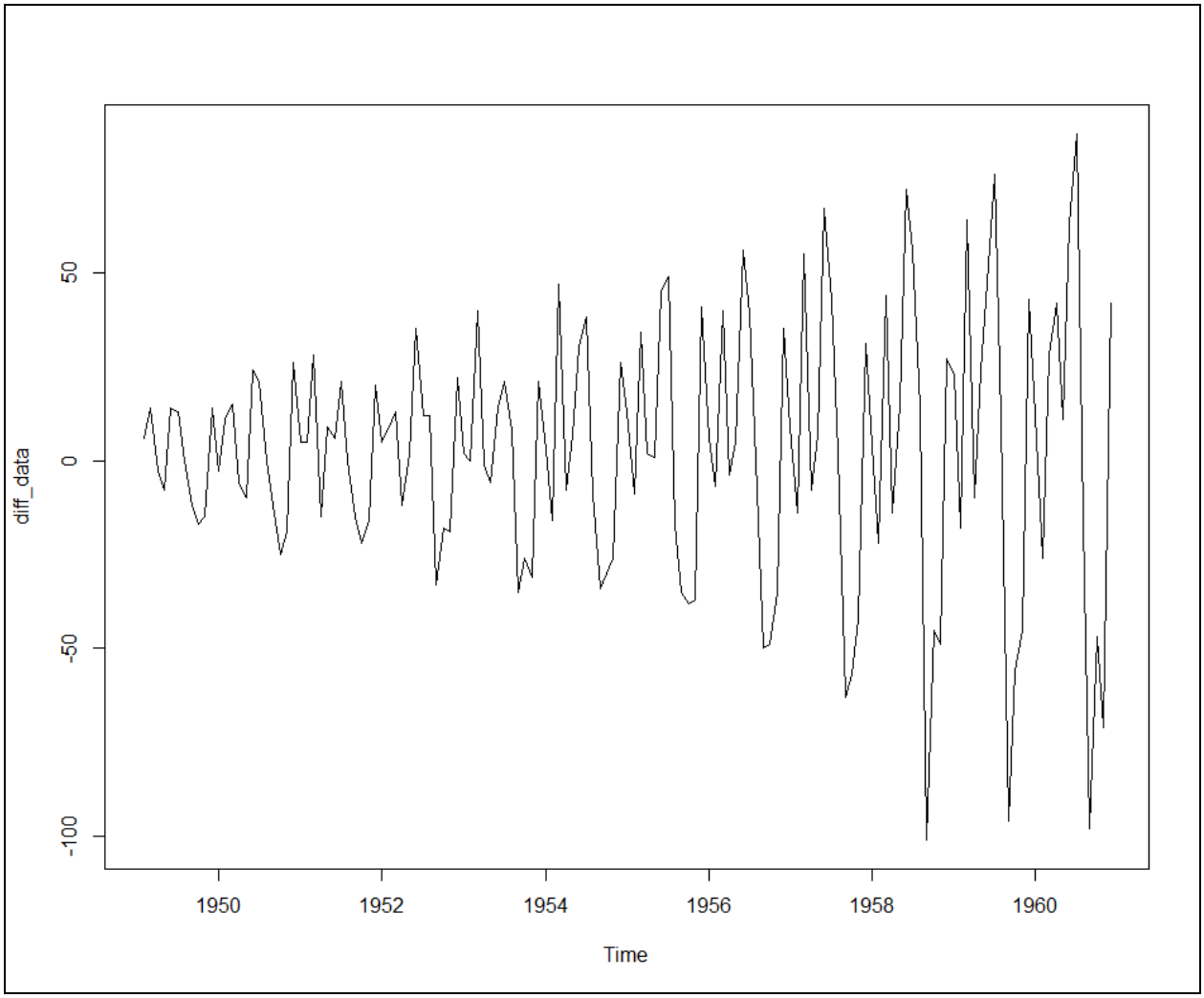
The output from the aov() function revealed some very interesting findings. The age factor shows a very low p-value, which is indicative of a statistically significant effect. The F value for age (39.89) further confirms this significance, indicating that the variation in exercise levels is substantially explained by age differences. These results strongly suggest that age plays a crucial role in influencing exercise levels, with a very small p-value underscoring this robustness.

2. Given that the AirPassengers data set has a substantial growth trend, use diff() to create a differenced data set. Use plot() to examine and interpret the results of differencing. Use cpt.var() to find the change point in the variability of the differenced time series. Plot the result and describe in your own words what the change point signifies.

The first step to completing this question is to library the changepoint package and use the diff() function to create a differenced data set. From there, we can create the first plot that shows the results of the differencing. The code that I used to accomplish this was as follows:

```
20  #Q2
21  library(changepoint)
22  diff_data <- diff(AirPassengers)
23  plot(diff_data)
```

And here is the resulting plot:

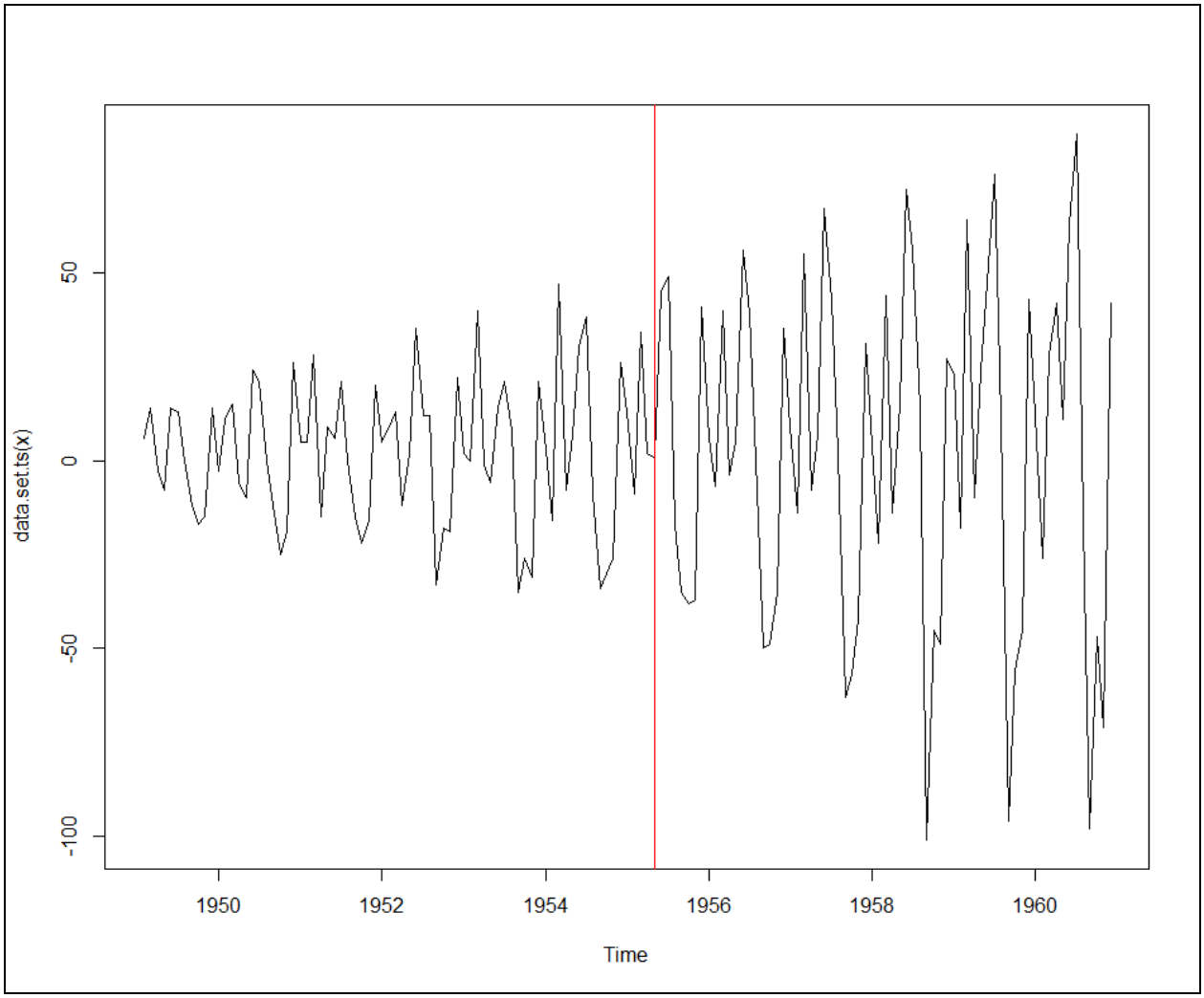


From this plot, we can observe the fact that the differenced data fluctuates around a constant mean rather than showing a clear upward or downward trend over time. The amplitude of the fluctuations seems to increase over time, suggesting that the variability of changes in passengers may be increasing as well. Lastly, there appears to be periodic peaks and troughs corresponding to the seasonal patterns in the data.

After this, we used the cpt.var() function to find the change point in the variability and then plot it again. To do this we can use a few simple lines of code:

```
24  cpt <- cpt.var(diff_data, method = "PELT")
25  plot(cpt)
26
```

With this, we get the resulting plot:



With this plot, we can see that the red vertical line appears to be around the year 1955. This suggests that the change point analysis has identified a shift in the variability of the time series around this time. Prior to the change point, the variability in the monthly differences of passenger numbers was relatively stable. However, after the change point, there is an apparent increase in the variability of these differences, indicating that the month-to-month changes in passenger numbers became more volatile. This could signify a change in the airline industry, changes in regulations, economic booms, or more advanced technology.

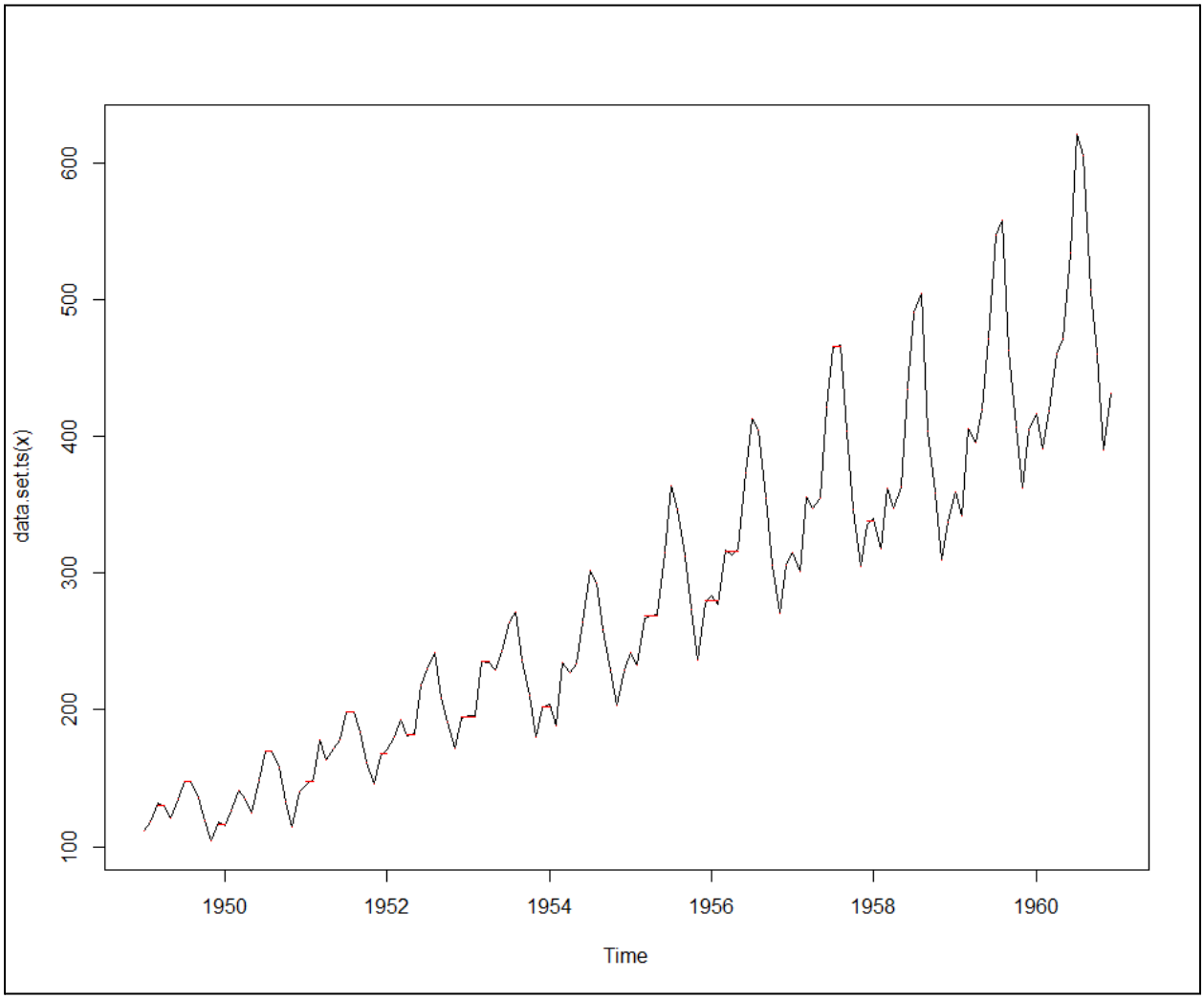
3. Use cpt.mean() on the AirPassengers time series. Plot and interpret the results. Compare the change point of the mean that you uncovered in this case to the change point in the variance that you uncovered in Exercise 5. What do these change points suggest about the history of air travel?

The code used for this question was overall quite simple because it was built off of the previous question. To plot the results, one simply had to use the cpt.mean() function from the AirPassengers data set. In total, here is the code that I used to plot the data:

```
#Q3
cpt_mean <- cpt.mean(AirPassengers, method = "PELT", class = TRUE)
plot(cpt_mean)

```

From this code above, I received this plot:



From this plot, it appears that the change point occurs around the early 1950s. This suggests that there was a notable increase in the mean number of air passengers around that time. Such a change could be attributed to a variety of factors including

technological advancements, economic factors like a post-war boom, and changes in social patterns with more people beginning to afford air travel.

Comparing this to the change point in the variance from the previous question, if the change in variance occurred at a different time than the change in mean, this could indicate that different factors influenced the number of passengers and the variability in those numbers. FOr instance, a change in mean might be associated with a steady increase in air travel popularity, whereas a change in variance might indicate periods of instability or rapid growth, which could have caused fluctuations in the number of passengers.

4. Find historical information about air travel on the internet and/or in reference materials that sheds light on the results from Exercises 5 and 6. Write a mini-article (less than 250 words) that interprets your statistical findings from Exercises 5 and 6 in the context of the historical information you found.

I just first want to say that I find this question extremely interesting, I have always been a casual fan of history and haven't studied it in a formal capacity since high school, so this was a very fun exercise for me! This aside here are some of the findings I made:

In the context of the post-war era, the 1950s saw a transformative period of commercial aviation. The decade saw technological advancements with the introduction of large, four-engine airliners like the Douglas DC-6 and the Boeing 377 Stratocruiser (cool name), which offered unprecedented comfort and efficiency. This period also heralded the rise of mass air travel in the United States. By 1955, more Americans were traveling by air than by train, signifying a shift in travel preferences and the beginning of an era where air travel would become more accessible to the general public.

The change points detected in the statistical analysis of the AirPassengers dataset reflect this historical shift. The significant change in the variance around the mid-1950s aligns with the industry's rapid growth and the onset of mass air travel. It indicates an increase in the variability of passenger numbers, likely a result of the growing popularity of air travel and the expanding capacity of airlines.

This interpretation of the statistical findings, when mapped onto the historical backdrop of the 1950s air travel boom, illustrates how data can reflect and reveal broader social and economic trends.

My source for all of this information comes from the National Air and Space Museum, in conjunction with Bill Steel for helping me find this information:

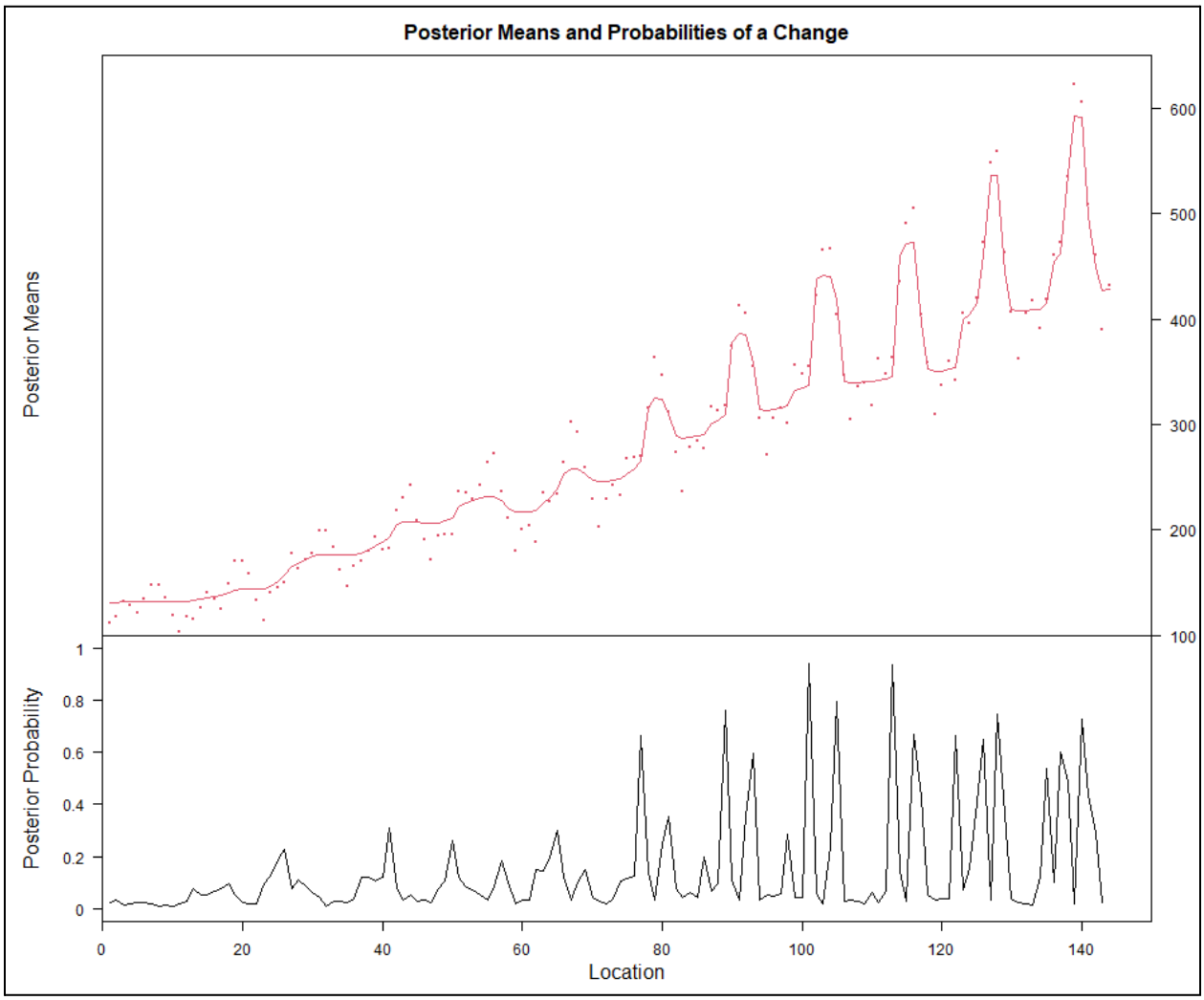
https://airandspace.si.edu/explore/stories/commercial-aviation-mid-century

5. Use bcp() on the AirPassengers time series. Plot and interpret the results. Make sure to contrast these results with those from Exercise 6.

In order to accomplish this task, I used the following code to install and library the package, call the bcp() function on the AirPassengers, and finally plot it:

```
#Q5
install.packages("bcp")
library(bcp)
bcp_result <- bcp(AirPassengers)
plot(bcp_result)
```

From this code I got the following result:



The BCP analysis provided some insightful visualizations into the subtle shifts and trends in air travel over time. The top graph delineates the estimated posterior means, reflecting the average number of passengers and how this average evolves. Notably, the spikes in the lower graph signify where there is a high probability of a shift in the average passenger count. These spikes reveal the possibility of multiple periods of change, suggesting (as we explored in the question above) a series of developments rather than singular events impacting air travel numbers.

Contrasting this with the results from Exercise 6, which identified clear and significant shifts in the mean, the BCP analysis offers a more nuanced perspective, indicating changes with varying levels of confidence. While the cpt.mean() function pointed to definitive moments of change, the BCP analysis suggests a gradual evolution, perhaps mirroring the gradual advances in aviation technology or incremental increases in the popularity of air travel.