ADM 2303 - Assignment 3

Renad Gharz

06/05/2022

1. Audio Streaming

An audio streaming service offers music and podcasts. Most music owned by record labels and the streaming service pays a copyright license fee each time a piece of music is streamed. By contrast, the streaming service purchases podcasts outright, paying a flat rate for each edition of the podcast, independent of the number of users that stream it. Forecasts of costs and revenues (from subscriptions and advertising) next year are Normally distributed and are given in the table as a central range within which there is 90% probability. For example, in units of $m, mean value of license fees = (24.5 + 27.2) / 2 = 25.85; P(24.5 < license fees < 27.2) = 0.9.

Cost of license fees (m) : $(24.5, 27.2) Cost of purchasing podcasts$(m): (1.1, 6.2) Revenues (from subscriptions and advertising) ($m):(43.9, 51.9)

First, we can create some vectors to make the analysis easier:

```
lc <- c(24.5,27.2) # Cost of license fees
pc <- c(1.1,6.2) # Cost of purchasing podcasts
rv <- c(43.9,51.9) # Revenues (subscriptions and advertising)
```

1.1 Coefficient of Variation

What is the coefficient of variation of each of the items in the table?

```
# Calculating the means of each item
mean_lc <- mean(lc) # Mean of license costs
mean_pc <- mean(pc) # Mean of podcasts costs
mean_rv <- mean(rv) # Mean of revenues

# Calculating the standard deviation of each item
sd_lc <- sd(lc) # Standard Deviation of license costs
sd_pc <- sd(pc) # Standard Deviation of podcasts costs
sd_rv <- sd(rv) # Standard Deviation of revenues

# Calculating the coefficients of variation of each item
coeff_var_lc <- sd_lc / mean_lc # Coefficient of variation of license costs
coeff_var_lc
```

```
## [1] 0.07385641
```

```
coeff_var_pc <- sd_pc / mean_pc # Coefficient of variation of podcast costs
coeff_var_pc
```

```
## [1] 0.9880122
```

```r
coeff_var_rv <- sd_rv / mean_rv # Coefficient of variation of revenues
coeff_var_rv
```

```
## [1] 0.1180972
```

1.2 Confidence Interval for Costs

> What is the central range within which there is 90% probability for total costs? State your assumption clearly and comment on whether you think it is valid. Are total costs Normally distributed? Give a reason for your answer.

```r
exp_tc <- mean_lc + mean_pc # Expected value of total costs
sd_tc <- sqrt(sd_lc^2 + sd_pc^2) # Standard deviation of total costs

z_score_tc <- qnorm(0.05) # Z score; Since this is a 2 sided 90% CI, we look for the Z-score of alpha/2

ci_lower_tc <- exp_tc + z_score_tc * sd_tc # CI lower bound
ci_upper_tc <- exp_tc - z_score_tc * sd_tc # CI upper bound
```

The central range within which $p = 0.9$ for total costs is $(22.788, 36.212)$ or $22.788 \leq x \leq 36.212$. The assumption made is that the two variables (LC and PC) are independent, and it isvalid because there is no logical reason to think that licensing costs for podcasts will affect the purchasing costs for podcasts and vice-versa.

1.3 Confidence Interval for Profits

> Revenues are related to the number of subscribers, as are licence fee costs, so that they are correlated. Revenues and total costs have a correlation coefficient of 0.58. What is the central range within which there is 90% probability for profits = revenues minus total costs?

```r
exp_pr <-  mean_rv - exp_tc # Expected value of profits
sd_pr <- sqrt(sd_rv^2 + sd_tc^2 - 2 * 0.58 * sd_rv * sd_tc) # Standard deviation of profits

z_score_pr <- qnorm(0.05) # Z score; Since this is a 2 sided 90% CI, we look for the Z-score of alpha/2

ci_lower_pr <- exp_pr + z_score_pr * sd_pr # CI lower bound
ci_upper_pr <- exp_pr - z_score_pr * sd_pr # CI upper bound
```

The central range within which $p = 0.9$ for profits is $(10.706, 26.094)$ or $10.706 \leq x \leq 26.094$.

1.4 Coefficient of Variation of Profits

> What is the coefficient of variation of profits?

```r
coeff_var_pr <- sd_pr / exp_pr
coeff_var_pr
```

```
## [1] 0.2541849
```

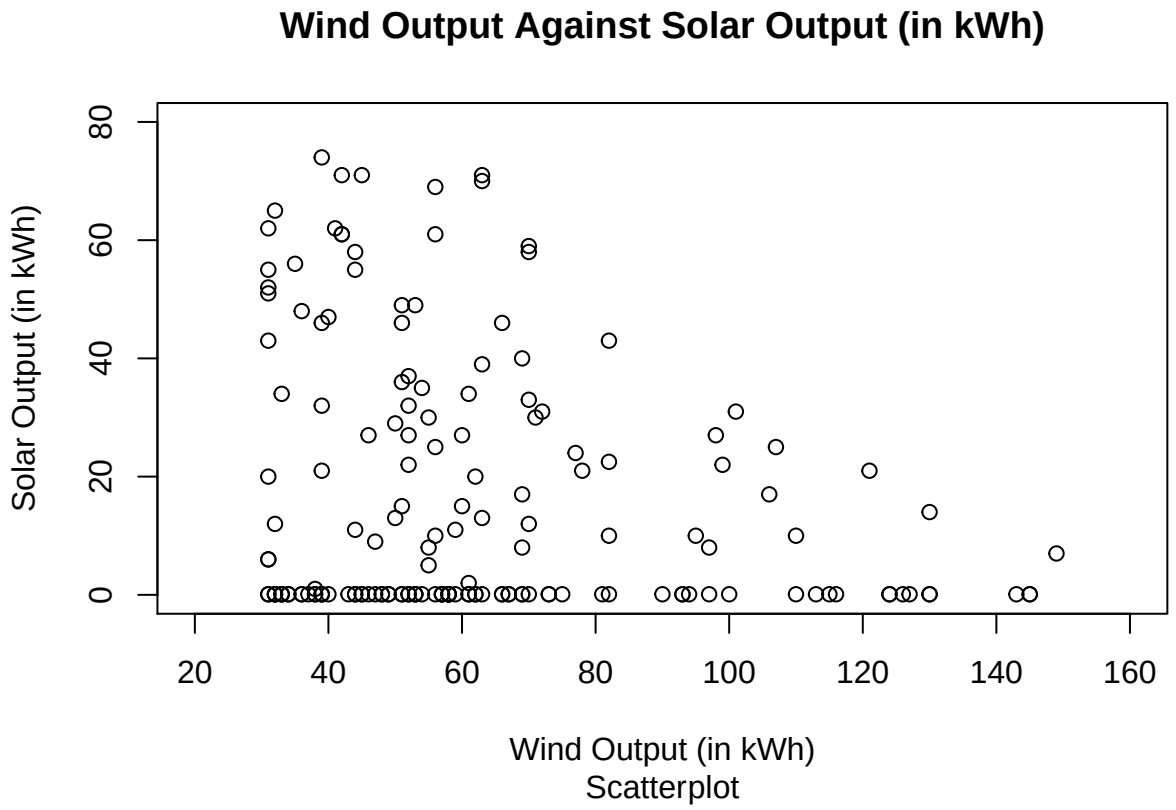
2. The Sun and the Wind

A company is planning to generate renewable electricity from an installation including both wind turbines and solar modules. It has selected a coastal site with plenty of wind and sun. The data file contains estimates of how much wind and solar electricity (kWh) would be generated each hour for a sample of one week from measurements of wind speed and solar radiation at the site.

2.1 Scatterplot

Draw a scatter diagram of solar output against wind output.

```
wind_sun_scatter <- plot(x = data_sun_wind$`Wind Output (kWh)`,
                         y = data_sun_wind$`Solar Output (kWh)`,
                         xlab = "Wind Output (in kWh)",
                         ylab = "Solar Output (in kWh)",
                         xlim = c(20,160),
                         ylim = c(0,80),
                         main = "Wind Output Against Solar Output (in kWh)",
                         sub = "Scatterplot") # Scatterplot of wind and sun data
```



2.2 Plot Analysis

From the diagram comment on the direction, form and strength of the relationship between solar and wind output.

Direction: negative Form: non-linear Strength: relatively weak

2.3 Extreme Points/Outliers

On the scatter diagram indicate 5 hours during which it was very windy at night. On the scatter diagram indicate 3 hours of exceptionally bright sunshine when it was calm (not very windy).

The three left-most and highest points are 3 hours (data points) with exceptionally bright sunshine, while the 5 right-most points near the x-axis are 5 hours (data points) where it was very windy at night.

2.4 Correlation Coefficient

Calculate the correlation coefficient between solar output and wind output.

```
sun_wind_cor <-
  cor(data_sun_wind$`Wind Output (kWh)`, data_sun_wind$`Solar Output (kWh)`)
# Correlation coefficient of wind and sun output
sun_wind_cor
```

```
## [1] -0.2308148
```

2.5 Negative Correlation Explanation

From your experience of sun and wind, explain why the correlation coefficient is negative.

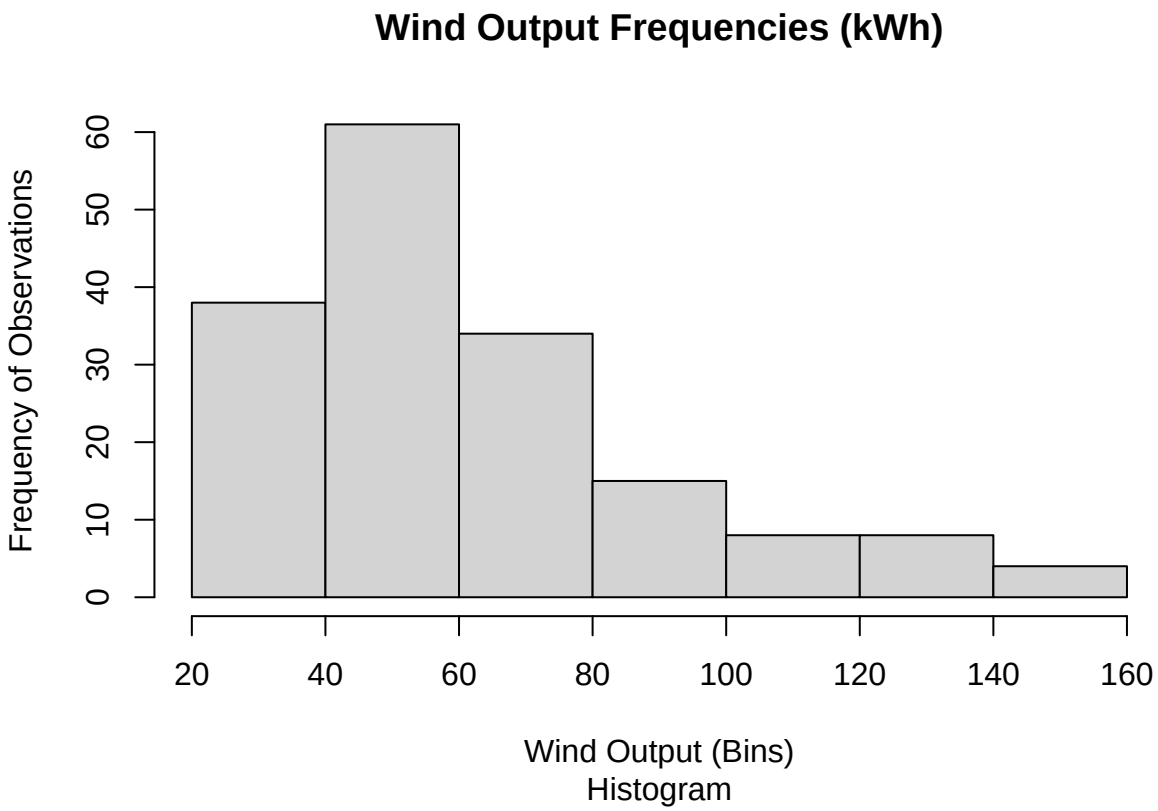
The correlation coefficient is negative because when it is very windy there is usually not much sunshine (it's either night time where there is no sunshine or it's extremely cloudy which blocks the sunshine) thus the less wind there is the more likely it is for there to be sunshine and vice-versa (more wind means less sunshine).

2.6 Histogram - Wind Output

Draw a histogram of wind output, giving the reason for choosing the number of bins/groups/bars.

```
wind_hist_bins <-
  round(log(168, 2), digits = 0)
# Log of 168 (number of observations in sample) with a base of 2 to find the initial number of bins

hist(data_sun_wind$`Wind Output (kWh)`,
     xlab = "Wind Output (Bins)",
     ylab = "Frequency of Observations",
     main = "Wind Output Frequencies (kWh)",
     sub = "Histogram",
     breaks = wind_hist_bins) # Histogram of wind output
```

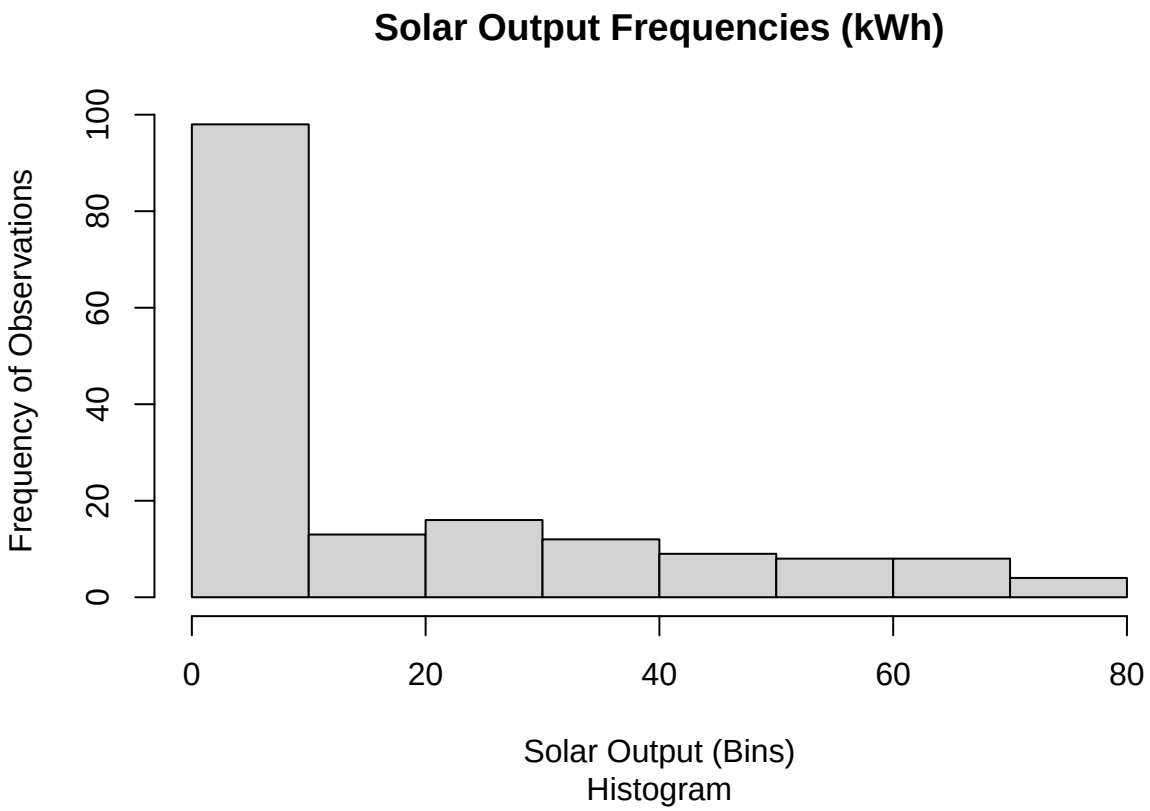


2.7

Draw a histogram of solar output, giving the reason for choosing the number of bins/groups/bars.

```
solar_hist_bins <-
  round(log(168, 2), digits = 0)
# Log of 168 (number of observations in sample) with a base of 2 to find the initial number of bins

hist(data_sun_wind$`Solar Output (kWh)`,
     xlab = "Solar Output (Bins)",
     ylab = "Frequency of Observations",
     main = "Solar Output Frequencies (kWh)",
     sub = "Histogram",
     breaks = solar_hist_bins) # Histogram of solar output
```



2.8 Scaterrplot vs Histogram

What can you tell from the scatterplot that you cannot tell from the histograms?

The scatter plot shows the linear relationship between the solar output and the wind output at the same time while the histogram only shows the grouped data for one variable at a time (we cannot see the relationship between the two using the histogram). The scatterplot also shows us all the data points which allows us to spot more effectively any outliers that may be present in the dataset and gain more information about the relationship of the 2 variables in the dataset. The histogram, on the other hand, does not show the individual data points, and only shows grouped data.