1.

Därför att städad data ger ett standardsätt att strukturera en “dataset” då hjälper denna för dataanalytiker att välja variabler som behövs för analys på lämplig och enkelt sätt.

I städad data observerar man:

- En kolumn som innehåll varje variabler.
- En rad som innehåll varje observation.
- Och en tabell som innehåll varje typ av observationsenhet.

Exempelvis, kan man anse den nedanstående två funktioner som används att städa data.

- filter (): välja eller ta bort observationer baserat på något tillstånd.
- arrange (): ändra ordningen (ökning eller minskning) av observationer.

2.

En dataframe är en tabell där varje kolumn innehåller värden för en variabel och varje rad innehåller en uppsättning värden från varje kolumn.

Dessutom har en dataframe följande egenskaper:

- Kolumnnamnen skall inte vara tomma.
- Radnamnen skall vara unika.
- Data som lagras i en dataram kan vara av numerisk, faktor- eller teckentyp.
- Varje kolumn bör innehålla samma antal dataobjekt.

Fördelarna med att använda dataframe är för att göra enklare datamanipulering, visualisering och modellering.

3.

En funktion är ett sätt med”statements”som organiserad tillsammans ska utföra en specifik uppgift.

Syftet med att skapa funktion är för att ha samma repetitiva kod som utför samma uppgift.

4.

I en for loop är antal iterationer (repetitioner) given från början.

I en while loop körs “statement” till ett visst villkor är uppfylld. Loopen bryt när “statement” blir FALSE (villkoren är inte uppfylld).

5.

```
if (test_expression) {statement }
```

```
# Om test_expression är TRUE, exekveras satsen. Men om det är FALSE händer ingenting.
```

Exempel : `x <- 5`

```
if(x > 0){print("counting number")}
```

6.

Ja, det innebär “coefficient of determination” som mäter hur närmare är datan från regression linjär.

OBS: R-squared är alltid mellan 0 and 1.

- R-squared = 0 betyder att modellen förklarar ingenting av variationen i data kring dess medelvärde.
- R-squared = 1 betyder att modellen förklarar alla variation av data kring dess medelvärde.

7.

Statistisk signifikans är en bestämning där resultaten i data inte bara kan förklaras av slumpmässiga faktorer. Dvs det är ett sätt att bestämma ett samband mellan två eller flera variabler beror på något annat än slumpen.

Statistisk signifikans används för att ge bevis för acceptera eller avvisa noll-hypotesen under hypotesprövning i genom att använda p-värde.

8.

När de slumpmässiga variablerna (X) i populationen är normala med medelvärde = μ och variation = σ^2.

9.

Konfidensintervall är ett sätt att mäta hur bra de samplade värde representerar populationen.

Den används att avgöra om uppskattningen av populationens medelvärde skulle falla någonstans i detta intervall eller inte. (CI: $\mu \pm \sigma$).

10.

Två variabler är korrelerade om den ena variabeln är beroende av den andra. Den betyder att någon ändring på den ena variabeln påverkar den andra när dem är korrelerade.

11.

En Outlier är en observation i en data som ligger långt från resten av observationerna. Dvs det är mycket större eller mindre än resten av värdena i data.

Vi kan hantera “outliers” I följande sätt:

- Att ta bort “outliers” från vår datan. (denna sätt är inte rekommenderas)

- Att utföra “quantile” baserad censurering. Dvs den lägsta “quantile” blir det minst “threshold” och det högsta “quantile” blir maximalt “threshold”.
- Att använda median istället för medelvärde eftersom medelvärdet är påverkat enormt på grund av “outliers”.

Referenser

1. Wickham, H., 2014. Tidy data. *Journal of statistical software*, *59*(1), pp.1-23.
2. https://www.tutorialspoint.com/r/r_data_frames.htm
3. R och statistik - Föreläsning 3.pdf
4. R och statistik - Föreläsning11&12.pdf
5. https://www.datamentor.io/r-programming/if-else-statement/
6. https://blog.minitab.com/en/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
7. https://www.investopedia.com/terms/s/statistical-significance.asp
8. https://www.investopedia.com/terms/s/statistically_significant.asp
9. https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/