Week 3 Assignment

Subject: Foundations of Statistical Analysis & Machine Learning (CSC1181)

Answers

Importing The simpsons_episodes.csv:

```
episodes <- read.csv("simpsons_episodes.csv", header = TRUE)
```

---

1. Statistical Analysis of IMDb Ratings

a) Calculating Mean Rating:

```
mean_rating <- mean(episodes$imdb_rating, na.rm = TRUE)
mean_rating
```

Output: 7.386097

---

b) Calculating the standard deviation:

```
sd_rating <- sd(episodes$imdb_rating, na.rm = TRUE)
sd_rating
```

Output: 0.7324394

---

c) Normal Distribution Curve of IMDB Ratings

```
min_rating <- min(episodes$imdb_rating, na.rm = TRUE)
max_rating <- max(episodes$imdb_rating, na.rm = TRUE)
curve(dnorm(x, mean=mean_rating, sd=sd_rating),
      from=min_rating-0.5, to=max_rating+0.5,
      col="blue", lwd=2,
      main="Normal Distribution of IMDb Ratings",
      xlab="IMDb Rating", ylab="Density")
hist(episodes$imdb_rating, breaks=20, freq=FALSE,
     col=rgb(0.8,0.8,0.8,0.4), add=TRUE)
abline(v=mean_rating, col="red", lwd=2, lty=2)
```

Output:

---

2. Probability Calculations

a) percentage of rating (imdb_rating) less or equal than 6:

```
prob_rating6 <- pnorm((6 - mean_rating)/sd_rating) * 100
paste0(round(prob_rating6, 2), "%")
```

Output: 2.92%

---

b) percentage of rating (imdb_rating) greater than 9:

```
prob_rating9plus <- (1 - pnorm((9 - mean_rating)/sd_rating)) * 100
paste0(round(prob_rating9plus, 2), "%")
```

Output: 1.38%

---

c) percentage of rating (imdb_rating) between 7 and 8:

```
prob_rating7to8 <- (pnorm((8 - mean_rating)/sd_rating) -
                    pnorm((7 - mean_rating)/sd_rating)) * 100
paste0(round(prob_rating7to8, 2), "%")
```

Output: 50%

---

d) percentage of viewers (us_viewers_in_millions) greater than 30:

```
mean_viewers <- mean(episodes$us_viewers_in_millions, na.rm = TRUE)
sd_viewers <- sd(episodes$us_viewers_in_millions, na.rm = TRUE)
prob_viewers30plus <- (1 - pnorm((30 - mean_viewers)/sd_viewers)) * 100
paste0(round(prob_viewers30plus, 2), "%")
```

Output: 0.23%

---

e) percentage of viewers(us_viewers_in_millions) less than 10:

```
prob_viewers10less <- pnorm((10 - mean_viewers)/sd_viewers) * 100
paste0(round(prob_viewers10less, 2), "%")
```

Output: 38.69%

---

f) percentage of viewers(us_viewers_in_millions) between 10 and 20:

```
prob_viewers10to20 <- (pnorm((20 - mean_viewers)/sd_viewers) -
                       pnorm((10 - mean_viewers)/sd_viewers)) * 100
paste0(round(prob_viewers10to20, 2), "%")
```

Output: 51.13%