**Student: Seif Kungulio**

**Date: 03/08/2025**

**Subject: Final Project**

**Class: DSCI 502**

**Section: 01W**

**Instructor: Sean Yang**

**File Name: ProjectFinal\_Kungulio\_Seif.docx**

1. Load the dataset day.csv into memory.

>

> ## 1. Load the dataset day.csv into memory.

>

> # Set the working directory to the correct location for the dataset.

> setwd("C:/PROJECTS/Maryville/DSCI-502/Week8")

>

> # Load necessary libraries for data manipulation, visualization, and reporting

> library(dplyr) # For data manipulation

> library(ggplot2) # For visualization

> library(knitr) # For creating formatted tables

>

> # Load the data from day.csv

> Bikes.df <- read.csv("day.csv")

>

> # Display the dimensions (rows and columns) of the dataframe

> dim(Bikes.df) # Shows the number of rows and columns in the dataset.

[1] 731 16

>

A screenshot of a computer program

AI-generated content may be incorrect.

> # Display column names of the data frame

> colnames(Bikes.df)

[1] "instant" "dteday" "season" "yr" "mnth" "holiday"

[7] "weekday" "workingday" "weathersit" "temp" "atemp" "hum"

[13] "windspeed" "casual" "registered" "cnt"

>

A screenshot of a computer screen

AI-generated content may be incorrect.

> # Display the structure of the data frame

> str(Bikes.df)

'data.frame': 731 obs. of 16 variables:

$ instant : int 1 2 3 4 5 6 7 8 9 10 ...

$ dteday : chr "2011-01-01" "2011-01-02" "2011-01-03" "2011-01-04" ...

$ season : int 1 1 1 1 1 1 1 1 1 1 ...

$ yr : int 0 0 0 0 0 0 0 0 0 0 ...

$ mnth : int 1 1 1 1 1 1 1 1 1 1 ...

$ holiday : int 0 0 0 0 0 0 0 0 0 0 ...

$ weekday : int 6 0 1 2 3 4 5 6 0 1 ...

$ workingday: int 0 0 1 1 1 1 1 0 0 1 ...

$ weathersit: int 2 2 1 1 1 1 2 2 1 1 ...

$ temp : num 0.344 0.363 0.196 0.2 0.227 ...

$ atemp : num 0.364 0.354 0.189 0.212 0.229 ...

$ hum : num 0.806 0.696 0.437 0.59 0.437 ...

$ windspeed : num 0.16 0.249 0.248 0.16 0.187 ...

$ casual : int 331 131 120 108 82 88 148 68 54 41 ...

$ registered: int 654 670 1229 1454 1518 1518 1362 891 768 1280 ...

$ cnt : int 985 801 1349 1562 1600 1606 1510 959 822 1321 ...

>

A screenshot of a computer screen

AI-generated content may be incorrect.

1. Perform the following data preparations using control structures:
2. Convert numerical season (1,2,3, 4) to characters (springer, summer, fall and winter)

>

> ## 2. Perform the following data preparations using control structures:

> ##### a. Convert numerical season (1,2,3, 4) to characters (springer, summer,

> ##### fall and winter)

>

> # Convert season to characters

> Bikes.df$season <- factor(Bikes.df$season,

+ levels = c(1, 2, 3, 4),

+ labels = c("spring", "summer", "fall", "winter"))

>

> # Verify conversion of 'season' variable

> str(Bikes.df$season)

Factor w/ 4 levels "spring","summer",..: 1 1 1 1 1 1 1 1 1 1 ...

>

A screenshot of a computer program

AI-generated content may be incorrect.

1. Convert numerical weathersit (1,2,3,4) to characters (Good, Mist, Bad, Severe)

>

> ##### b. Convert numerical weathersit (1,2,3,4) to characters (Good, Mist, Bad,

> ##### Severe)

>

> # Convert weathersit to characters

> Bikes.df$weathersit <- factor(Bikes.df$weathersit,

+ levels = c(1, 2, 3, 4),

+ labels = c("Good", "Mist", "Bad", "Severe"))

>

> # Verify conversion of 'weathersit' variable

> str(Bikes.df$weathersit)

Factor w/ 4 levels "Good","Mist",..: 2 2 1 1 1 1 2 2 1 1 ...

>

A screenshot of a computer program

AI-generated content may be incorrect.

1. Consider the following predictors, season, holiday, workingday, weathersit, atemp, hum, windspeed, casual and list all categorical variables from this list and convert them to factors.

>

> ## 3. Consider the following predictors, season, holiday, workingday,

> ## weathersit, atemp, hum, windspeed, casual and List all categorical

> ## variables from this list and convert them to factors

>

> # Convert categorical variables to factors

> categorical\_vars <- c("season", "holiday", "workingday", "weathersit")

> Bikes.df[categorical\_vars] <- lapply(Bikes.df[categorical\_vars], as.factor)

>

> # Display the statistical summary for factor variables

> summary(Bikes.df[categorical\_vars])

season holiday workingday weathersit

spring:181 0:710 0:231 Good :463

summer:184 1: 21 1:500 Mist :247

fall :188 Bad : 21

winter:178 Severe: 0

>

A screenshot of a computer program

AI-generated content may be incorrect.

1. Calculate the minimum, maximum, mean, median, standard deviation and three quartiles (25th, 50th and 75th percentiles) of cnt.

>

> ## 4. Calculate the minimum, maximum, mean, median, standard deviation and

> ## three quartiles (25th, 50th and 75th percentiles) of cnt.

>

> # Summary Statistics for cnt

> cnt\_stats <- Bikes.df %>%

+ summarise(minimum = min(cnt), maximum = max(cnt), mean = mean(cnt),

+ median = median(cnt), sd = sd(cnt), Q1 = quantile(cnt, 0.25),

+ Q2 = quantile(cnt, 0.5), Q3 = quantile(cnt, 0.75))

> # Display summary statistics in a table

> kable(cnt\_stats)

| minimum| maximum| mean| median| sd| Q1| Q2| Q3|

|------------:|-------------:|---------:|---------:|-------:|----:|-----:|-----:|

| 22| 8714| 4504.349| 4548| 1937.211| 3152| 4548| 5956|

>

A computer screen with text and numbers

AI-generated content may be incorrect.

1. Calculate the minimum, maximum, mean, median, standard deviation and three quartiles (25th, 50th and 75th percentiles) of registered.

>

> ## 5. Calculate the minimum, maximum, mean, median, standard deviation and

> ## three quartiles (25th, 50th and 75th percentiles) of registered.

>

> # Summary Statistics for registered users

> registered\_stats <- Bikes.df %>%

+ summarise(minimum = min(registered), maximum = max(registered),

+ mean = mean(registered), median = median(registered),

+ sd = sd(registered), Q1 = quantile(registered, 0.25),

+ Q2 = quantile(registered, 0.5), Q3 = quantile(registered, 0.75))

>

> # Display summary statistics in a table

> kable(registered\_stats)

| minimum| maximum| mean| median| sd| Q1| Q2| Q3|

|-------:|-------:|--------:|------:|--------:|----:|----:|------:|

| 20| 6946| 3656.172| 3662| 1560.256| 2497| 3662| 4776.5|

>

A computer screen with white text

AI-generated content may be incorrect.

1. Calculate the correlation coefficient of the two variables: registered and cnt. Do they have a strong relationship?

>

> ## 6. Calculate the correlation coefficient of the two variables: registered

> ## and cnt. Do they have a strong relationship?

>

> # Correlation between registered and cnt

> correlation <- cor(Bikes.df$registered, Bikes.df$cnt)

>

> # Print correlation result

> print(correlation)

[1] 0.9455169

>

A computer screen shot of a black screen

AI-generated content may be incorrect.

The correlation coefficient between "registered" and "cnt" is **0.9455**, which indicates a **very strong positive** relationship between the two variables. A correlation coefficient (**r**) close to **1** means that as the number of registered users increases, the total bike count ("cnt") also increases significantly. This suggests that registered users make up a large portion of the total bike rentals.

1. Calculate the frequency table of season? What’s the mode of season variable?

>

> ## 7. Calculate the frequency table of season? What’s the mode of

> ## season variable?

>

> # Frequency table of 'season' and identify the most common season

> season\_freq <- table(Bikes.df$season)

> print(season\_freq)

spring summer fall winter

181 184 188 178

> mode\_season <- names(which.max(season\_freq))

> print(paste("Mode of season: ", mode\_season))

[1] "Mode of season: fall"

>

A screen shot of a computer screen

AI-generated content may be incorrect.

1. Calculate the cross table of season and weathersit, then produce proportions by rows and columns respectively.

>

> ## 8. Calculate the cross table of season and weathersit, then produce

> ## proportions by rows and columns respectively.

>

> # Create a cross table between 'season' and 'weathersit' and compute proportions

> season\_weather\_table <- table(Bikes.df$season, Bikes.df$weathersit)

> season\_weather\_row\_prop <- prop.table(season\_weather\_table, 1)

> season\_weather\_col\_prop <- prop.table(season\_weather\_table, 2)

> kable(season\_weather\_table)

| | Good| Mist| Bad| Severe|

|:------|----:|----:|---:|------:|

|spring | 111| 66| 4| 0|

|summer | 113| 68| 3| 0|

|fall | 136| 48| 4| 0|

|winter | 103| 65| 10| 0|

> kable(season\_weather\_row\_prop)

| | Good| Mist| Bad| Severe|

|:------|---------:|---------:|---------:|------:|

|spring | 0.6132597| 0.3646409| 0.0220994| 0|

|summer | 0.6141304| 0.3695652| 0.0163043| 0|

|fall | 0.7234043| 0.2553191| 0.0212766| 0|

|winter | 0.5786517| 0.3651685| 0.0561798| 0|

> kable(season\_weather\_col\_prop)

| | Good| Mist| Bad| Severe|

|:------|---------:|---------:|---------:|------:|

|spring | 0.2397408| 0.2672065| 0.1904762| NaN|

|summer | 0.2440605| 0.2753036| 0.1428571| NaN|

|fall | 0.2937365| 0.1943320| 0.1904762| NaN|

|winter | 0.2224622| 0.2631579| 0.4761905| NaN|

>

A screenshot of a computer program

AI-generated content may be incorrect.

1. Please plot the histogram and density of the cnt and add the vertical line denoting the mean using ggplot2.

>

> ## 9. Please plot the histogram and density of the cnt and add the vertical

> ## line denoting the mean using ggplot2.

>

> # Plot histogram and density of 'cnt' with mean line

> ggplot(Bikes.df, aes(x = cnt)) +

+ geom\_histogram(binwidth = 500, fill = "green") +

+ geom\_density(color = "red", lwd = 2) +

+ geom\_vline(aes(xintercept = mean(cnt)),

+ color = "black", linetype = "dashed", lwd = 1) +

+ ggtitle("Histogram of cnt") + theme\_test()

>

A green and red graph

AI-generated content may be incorrect.

1. Please scatter plot of cnt (y-axis) against registered (x-axis) and add the trend line using ggplot2.

>

> ## 10. Please scatter plot of cnt (y-axis) against registered (x-axis) and add

> ## the trend line using ggplot2.

>

> # Scatter plot of 'cnt' vs 'registered' with trend line

> ggplot(Bikes.df, aes(x = registered, y = cnt)) +

+ geom\_point() +

+ geom\_smooth(method = "lm", col = "red") +

+ theme\_test() +

+ ggtitle("Scatter plot of cnt vs registered")

`geom\_smooth()` using formula = 'y ~ x'

>

A black and red line

AI-generated content may be incorrect.

1. Please plot the barplot of season and weathersit on the same barplot using ggplot2

>

> ## 11. Please plot the barplot of season and weathersit on the same barplot

> ## using ggplot2

>

> # Barplot of season and weathersit

> ggplot(Bikes.df, aes(x = season, fill = weathersit)) +

+ geom\_bar(position = "dodge") +

+ scale\_fill\_manual(values=c("#008000", "#0000FF", "#FF0000")) +

+ ggtitle("Barplot of season and weathersit") + theme\_test()

>

A graph of different colored bars

AI-generated content may be incorrect.

1. Please boxplot cnt (y-axis) against weathersit (x-axis) and save the graph in a file, cntweather.jpg, using ggplot2. Are there any differences in cnt with respect to weathersit?

>

> ## 12. Please boxplot cnt (y-axis) against weathersit (x-axis) and save the

> ## graph in a file, cntweather.jpg, using ggplot2. Are there any differences

> ## in cnt with respect to weathersit?

>

> # Box plot of 'cnt' vs 'weathersit', saved as 'cntweather.jpg'

> cnt\_weather\_boxplot <- ggplot(Bikes.df,

+ aes(x = weathersit,

+ y = cnt,

+ fill = weathersit)) +

+ geom\_boxplot() + theme\_test() +

+ scale\_fill\_manual(values = c("#008000", "#0000FF", "#FF0000")) +

+ ggtitle("Boxplot cnt vs weathersit")

> plot(cnt\_weather\_boxplot)

> ggsave("cntweather.jpg")

Saving 12.1 x 5.96 in image

>

A diagram of a graph

AI-generated content may be incorrect.

1. Build the following multiple linear regression models:
2. Perform multiple linear regression with cnt as the response and the predictors are: season, weathersit, atemp, and registered. Write down the math formula with numerical coefficients for predictors atemp and registered and skip the coefficients for season and weathersit.

>

> ## 13. Build the following multiple linear regression models:

> ###### a. Perform multiple linear regression with cnt as the response and the

> ###### predictors are: season, weathersit, atemp, and registered.

> ###### Write down the math formula with numerical coefficients for

> ###### predictors atemp and registered and skip the coefficients

> ###### for season and weathersit.

>

> # Model 1

> model1 <- lm(cnt ~ season + weathersit + atemp +

+ registered, data = Bikes.df)

> summary(model1)

Call:

lm(formula = cnt ~ season + weathersit + atemp + registered,

data = Bikes.df)

Residuals:

Min 1Q Median 3Q Max

-967.5 -359.7 -145.7 142.0 2525.1

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -132.23314 81.74777 -1.618 0.1062

seasonsummer 354.90830 76.22493 4.656 3.84e-06 \*\*\*

seasonfall 191.56343 97.30364 1.969 0.0494 \*

seasonwinter 170.78460 67.69057 2.523 0.0118 \*

weathersitMist -201.24823 44.41700 -4.531 6.87e-06 \*\*\*

weathersitBad -579.16760 129.65059 -4.467 9.20e-06 \*\*\*

atemp 1657.84472 230.71955 7.186 1.67e-12 \*\*\*

registered 1.02693 0.01726 59.481 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 551.8 on 723 degrees of freedom

Multiple R-squared: 0.9196, Adjusted R-squared: 0.9189

F-statistic: 1182 on 7 and 723 DF, p-value: < 2.2e-16

>

**Math formula:**

**cnt = -132.23314 + (1657.84472\*atemp) + (1.02693\*registered)**

A screenshot of a computer program

AI-generated content may be incorrect.

1. Preform multiple linear regression with cnt as the response and the predictors are: season, workingday, weathersit, atemp, and registered. Write down the math formula with numerical coefficients for predictors atemp and registered and skip the coefficients for season, workingday and weathersit.

>

> ###### b. Preform multiple linear regression with cnt as the response and the

> ###### predictors are: season, workingday, weathersit, atemp, and

> ###### registered. Write down the math formula with numerical

> ###### coefficients for predictors atemp and registered and skip the

> ###### coefficients for season, workingday, and weathersit.

>

> # Model 2

> model2 <- lm(cnt ~ season + workingday + weathersit +

+ atemp + registered, data = Bikes.df)

> summary(model2)

Call:

lm(formula = cnt ~ season + workingday + weathersit + atemp +

registered, data = Bikes.df)

Residuals:

Min 1Q Median 3Q Max

-952.75 -219.47 -44.07 181.28 1613.56

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 234.4625 55.5376 4.222 2.73e-05 \*\*\*

seasonsummer 237.1732 50.6951 4.678 3.45e-06 \*\*\*

seasonfall 56.2383 64.6781 0.870 0.38486

seasonwinter -33.6740 45.3894 -0.742 0.45839

workingday1 -961.2252 31.6532 -30.367 < 2e-16 \*\*\*

weathersitMist -82.9332 29.7106 -2.791 0.00539 \*\*

weathersitBad -154.5223 87.1043 -1.774 0.07649 .

atemp 1287.9916 153.4798 8.392 2.51e-16 \*\*\*

registered 1.1714 0.0124 94.484 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 365.9 on 722 degrees of freedom

Multiple R-squared: 0.9647, Adjusted R-squared: 0.9643

F-statistic: 2467 on 8 and 722 DF, p-value: < 2.2e-16

>

**Math formula:**

**cnt = 234.4625 + (1287.9916\*atemp) + (1.1714\*registered)**

A screenshot of a computer program

AI-generated content may be incorrect.

1. Preform multiple linear regression with cnt as the response and the predictors are: season, holiday, workingday, weathersit, atemp, hum, windspeed, and registered. Write down the math formula with numerical coefficients for predictors atemp, hum, windspeed, and registered and skip the coefficients for season, holiday, workingday and weathersit.

>

> ###### c. Preform multiple linear regression with cnt as the response and the

> ###### predictors are: season, holiday, workingday, weathersit, atemp, hum,

> ###### windspeed, and registered. Write down the math formula with numerical

> ###### coefficients for predictors atemp, hum, windspeed, and registered

> ###### and skip the coefficients for season, holiday,

> ###### workingday and weathersit.

>

> # Model 3

> model3 <- lm(cnt ~ season + holiday + workingday + weathersit +

+ atemp + hum + windspeed + registered, data = Bikes.df)

> summary(model3)

Call:

lm(formula = cnt ~ season + holiday + workingday + weathersit +

atemp + hum + windspeed + registered, data = Bikes.df)

Residuals:

Min 1Q Median 3Q Max

-845.72 -228.33 -35.32 171.66 1611.31

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 580.0198 101.9434 5.690 1.85e-08 \*\*\*

seasonsummer 239.8855 50.0548 4.792 2.00e-06 \*\*\*

seasonfall 39.9719 63.8698 0.626 0.53162

seasonwinter -26.5911 45.5651 -0.584 0.55968

holiday1 -245.4140 82.7552 -2.966 0.00312 \*\*

workingday1 -977.0324 32.0082 -30.524 < 2e-16 \*\*\*

weathersitMist -32.9389 35.6562 -0.924 0.35590

weathersitBad -44.8026 93.3135 -0.480 0.63128

atemp 1378.9261 156.8044 8.794 < 2e-16 \*\*\*

hum -363.9911 130.9036 -2.781 0.00557 \*\*

windspeed -637.7530 189.6037 -3.364 0.00081 \*\*\*

registered 1.1607 0.0125 92.872 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 360.8 on 719 degrees of freedom

Multiple R-squared: 0.9658, Adjusted R-squared: 0.9653

F-statistic: 1848 on 11 and 719 DF, p-value: < 2.2e-16

>

**Math formula:**

**cnt = 580.0198 + (1378.9261 \* atemp) + (1.1607 \* registered) + (-637.7530 \* windspeed) + (-363.9911 \* hum)**

A screenshot of a computer program

AI-generated content may be incorrect.

1. Which model do you recommend to the management based on adjusted R squared? Justify your answer.

>

> ###### d. Which model do you recommend to the management based on

> ###### adjusted R squared? Justify your answer.

>

> # Select best regression model based on adjusted R-squared

> adjusted\_r\_squared <- c(summary(model1)$adj.r.squared,

+ summary(model2)$adj.r.squared,

+ summary(model3)$adj.r.squared)

>

> best\_model <- which.max(adjusted\_r\_squared)

> paste("Best model based on Adjusted R-squared: Model", best\_model)

[1] "Best model based on Adjusted R-squared: Model 3"

>

A screenshot of a computer error

AI-generated content may be incorrect.

1. Summarize Question 13-C using R markdown to generate a reproducible report. Include the following scripts in your R markdown file:
2. Load the data as specified in Question 1.
3. Convert the two variables as specified in Question 2.
4. Convert the categorical variables to factors as specified in Question 3
5. Build a linear model as specified in Question 13-C. Use R markdown to report the math formula with numerical coefficients for predictors atemp, hum, windspeed, and registered. Skip the coefficients for season, holiday, workingday and weathersit.
6. Build the following logistic models:
7. forecast holiday using cnt, season, and registered.

>

> ## 15. Build the following logistic models:

>

> # McFadden/pseudo R squared

> pseudo\_r2 <- function(model) { 1 - (model$deviance / model$null.deviance) }

>

> ###### a. forecast holiday using cnt, season, and registered.

> logit1 <- glm(holiday ~ cnt + season + registered,

+ data = Bikes.df, family = binomial())

> pseudo\_r2(logit1)

[1] 0.1070295

>

A screenshot of a computer program

AI-generated content may be incorrect.

1. forecast the holiday using cnt, season, weathersit, and registered

>

> ###### b. forecast the holiday using cnt, season, weathersit, and registered

> logit2 <- glm(holiday ~ cnt + season + weathersit + registered,

+ data = Bikes.df, family = binomial())

> pseudo\_r2(logit2)

[1] 0.123026

>

A black screen with white text

AI-generated content may be incorrect.

1. forecast the holiday using cnt, season, weathersit, workingday, and registered

>

> ###### c. forecast the holiday using cnt, season, weathersit, workingday,

> ###### and registered

> logit3 <- glm(holiday ~ cnt + season + weathersit + workingday + registered,

+ data = Bikes.df, family = binomial())

Warning message:

glm.fit: fitted probabilities numerically 0 or 1 occurred

> pseudo\_r2(logit3)

[1] 0.2888301

>

A black screen with white text

AI-generated content may be incorrect.

1. Which model do you recommend to the management based on McFadden/pseudo-R squared to? Justify your answer

>

> ###### d. Which model do you recommend to the management based on

> ###### McFadden/pseudo R squared to? Justify your answer

>

> # Select best logistic model based on McFadden's pseudo R-squared

> mcfadden\_r2 <- c(pseudo\_r2(logit1), pseudo\_r2(logit2), pseudo\_r2(logit3))

> best\_logit\_model <- which.max(mcfadden\_r2)

> paste("Best logistic model based on McFadden R-squared: Model", best\_logit\_model)

[1] "Best logistic model based on McFadden R-squared: Model 3"

>

A screenshot of a computer error

AI-generated content may be incorrect.