Final take-home exam

Pooja Sadarangani

2022-12-08

#1 Data exploration and multiple regression (90pt)

1.1 Explore life expectancy (50pt)

1. (2pt) Explain what is life expectancy. Here we talk about period life expectancy at birth, not cohort life expectancy.

LEFT Ans. Life expectancy is a statistical measure of how long an organism is expected to live based on its birth year and other factors such as age and gender.

2. (6pt) Load and clean the data–remove all cases with missing life expectancy, year and country name or code. You may have to return here later to improve cleaning if you discover more issues below. How many good cases do we have?Explain what steps you do, and in case the task is ambiguous, explain why you take exactly these steps too.

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v stringr 1.4.1
## v tidyr   1.2.1     v forcats 0.5.2
## v readr   2.1.3
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
## [1] 13055    25
```

```
## [1] 13055
```

```
##  [1] "iso3"               "name"               "iso2"
##  [4] "region"             "sub.region"         "intermediate.region"
##  [7] "time"               "totalPopulation"    "fertilityRate"
## [10] "lifeExpectancy"     "childMortality"     "youthFemaleLiteracy"
## [13] "youthMaleLiteracy"  "adultLiteracy"      "GDP_PC"
## [16] "accessElectricity"  "agriculturalLand"   "agricultureTractors"
## [19] "cerealProduction"   "fertilizerHa"       "co2"
## [22] "greenhouseGases"    "co2_PC"             "pm2.5_35"
## [25] "battleDeaths"
```

```
##  [1] "iso3"                "name"                "iso2"
##  [4] "region"              "sub.region"          "intermediate.region"
##  [7] "year"                "totalPopulation"     "fertilityRate"
## [10] "lifeExpectancy"      "childMortality"      "youthFemaleLiteracy"
## [13] "youthMaleLiteracy"   "adultLiteracy"       "GDP_PC"
## [16] "accessElectricity"   "agriculturalLand"    "agricultureTractors"
## [19] "cerealProduction"    "fertilizerHa"        "co2"
## [22] "greenhouseGases"     "co2_PC"              "pm2.5_35"
## [25] "battleDeaths"
```

```
## The total number of NA values in the dataset are: 103406

## The total number of NA values in column lifeExpectancy are: 1325

## The total number of NA values in column iso2 are: 0

## The total number of NA values in column iso3 are: 0

## The total number of NA values in column name are: 0

## The total number of NA values in column name are: 36

## [1] TRUE

## The total number of NA values in column lifeExpectancy are: 0

## The total number of NA values in column iso2 are: 0

## The total number of NA values in column iso3 are: 0

## The total number of NA values in column name are: 0

## The total number of NA values in column name are: 0

## The total number of good cases after cleaning the data is  11618
```

3. (6pt) Now it is time to do some brief exploration:

###(a) How many countries do we have in these data? ###(b) What is the first and last year with valid life expectancy data? ###(c) What is the lowest and highest life expectancy values? Which country/year do they correspond to? ###(d) If you did this correctly, you see that the shortest life expectancy corresponds to a well-known event. What is the event? (You may consult wikipedia if you do not know).

(a) Number of countries in the dataset

```
## The number of countries in the dataset are  204
```

(b) First and Last year with valid expectancy

```
## First year with valid life expectancy is  1960

## Last year with valid life expectancy is  2019
```

(c) Lowest and Highest Life Expectancy with their corresponding countries and year

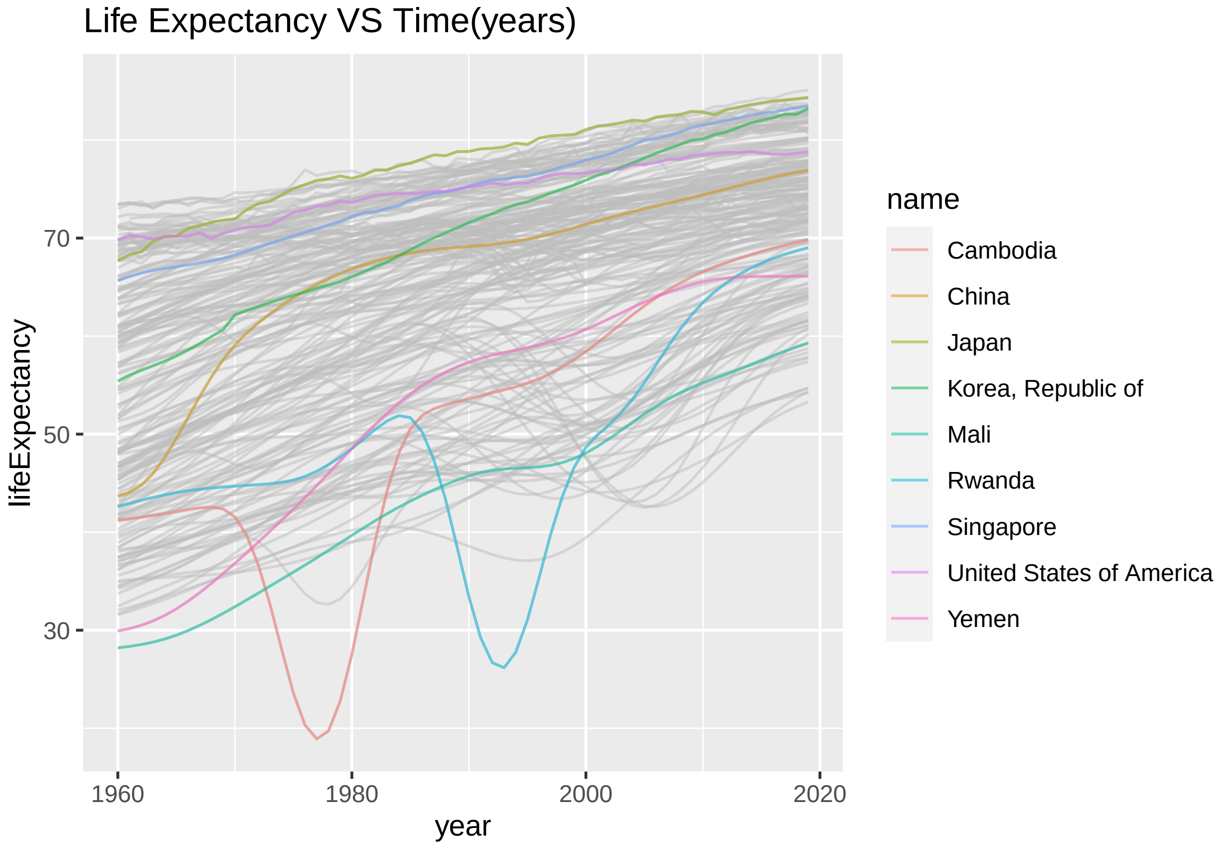
```
## Lowest Life Expectancy is  18.907 and corresponding country is  Cambodia  and corresponding year is

## Highest Life Expectancy is  85.41707 and corresponding country is  San Marino  and corresponding yea
```

(d) The shortest life expectancy corresponds to the Cambodia Genocide event.

4. (10pt) Next, lets plot the life expectancy over time for all countries (there are many of them).

Make a plot where you show life expectancy in each country versus time. Highlight the U.S., South Korea, Cambodia, and China on this graph.

Choose yourself a few additional countries, and explain why do you think it is interested to look at those countries.



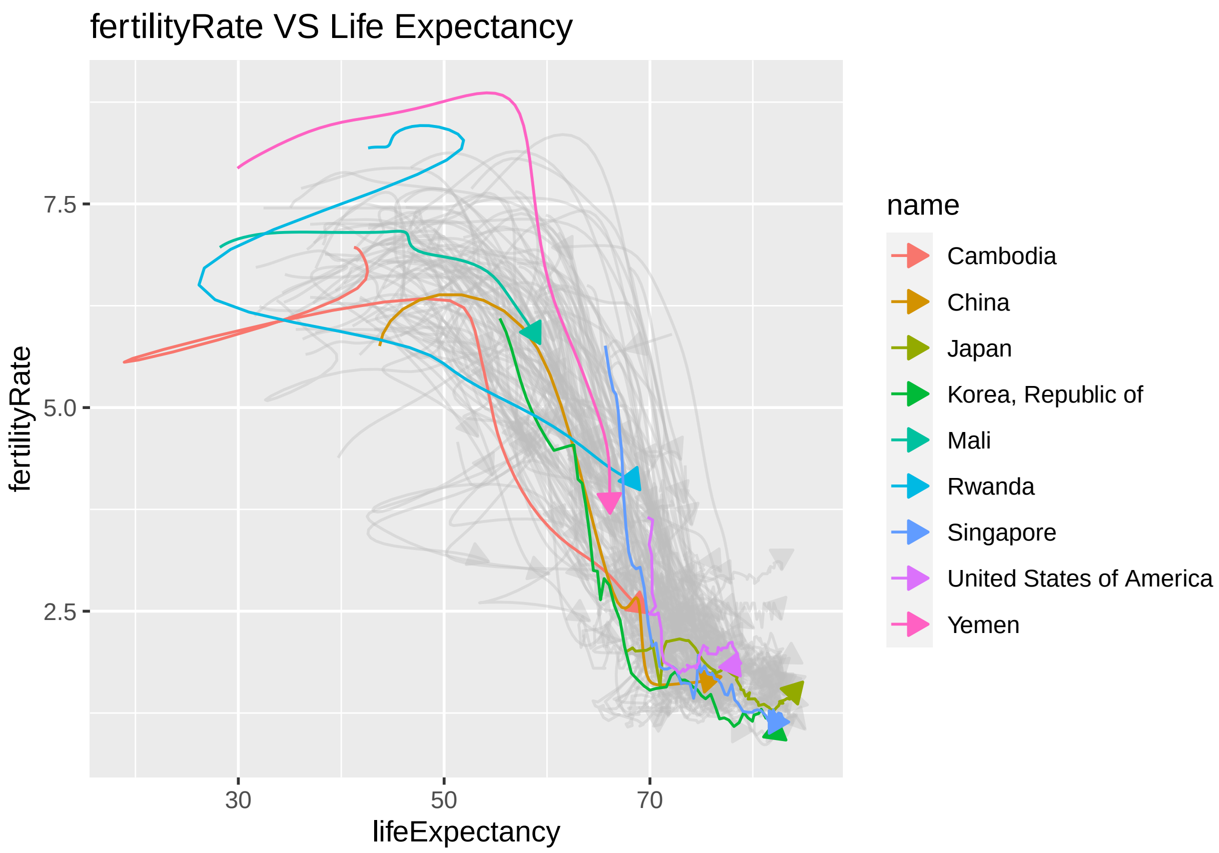
### I have chosen Mali, Yemen, and Rwanda as additional countries as all three have them had really low life expectancy (<30) at some point of time. ALong with this, I have also chosen Japan and Singapore as they had really high life expectancy at some point of time (>83)

5. (8pt) Explain what do you see on the graph. What is the overall picture? How do the selected countries behave? Anything else interesting you see?

Ans. Overall, the life expectancy has increased for all countries from 1960 to 2020. However, Cambodia and Rwanda had major drops in their life expectancy and today, they still have low life expectance as compared to countries like Japan. Japan, US, and Singapore have consistantly had high life expectancy, which seems to be growing with time.

6. (10pt) Now, let's look at how are life expectancy and fertility related. Make a fertility rate versus life expectancy plot of all countries with selected countries highlighted. Use arrows to mark which way the time goes on the figure.

```
## Warning: Removed 18 row(s) containing missing values (geom_path).
```

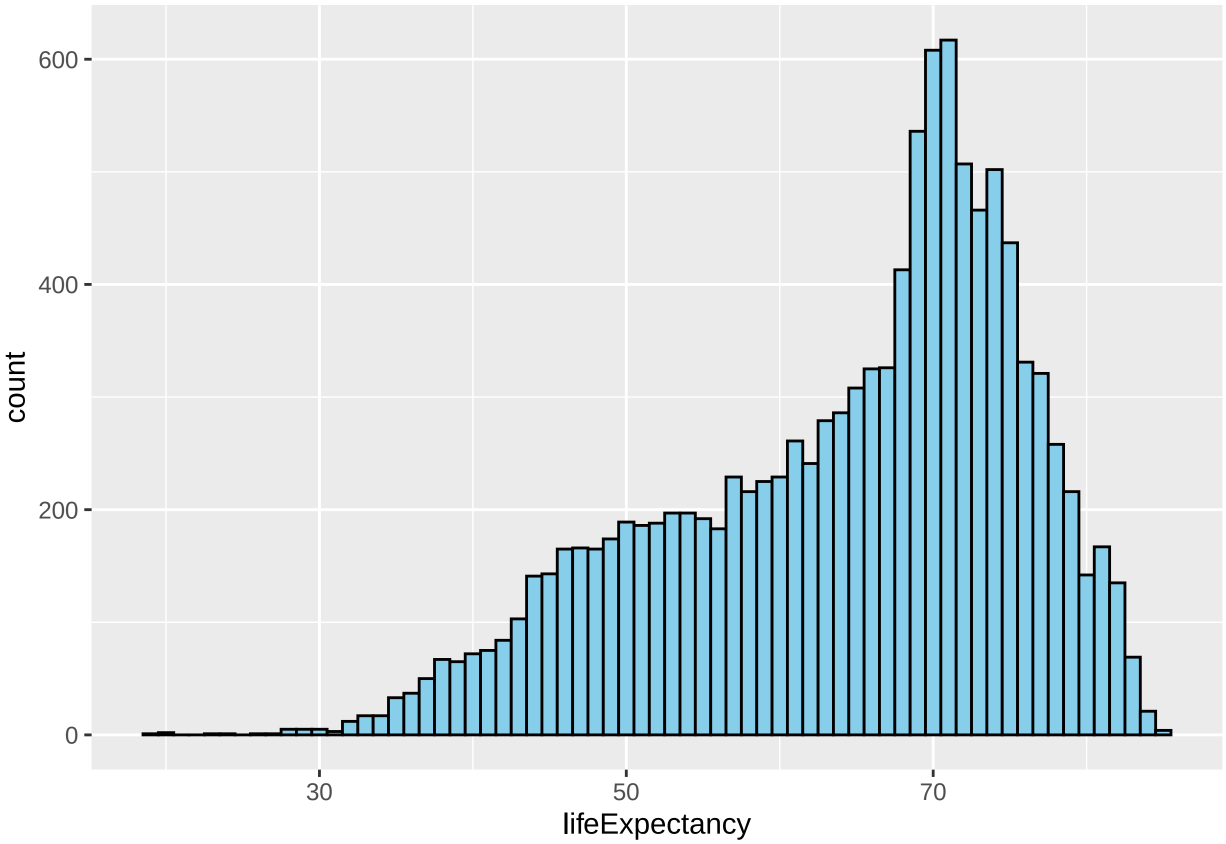

fertilityRate VS Life Expectancy
fertilityRate
7.5
5.0
2.5
lifeExpectancy
30
50
70
name
Cambodia
China
Japan
Korea, Republic of
Mali
Rwanda
Singapore
United States of America
Yemen

7. (8pt) Comment the results. Where is the world going? Where are the highlighted countries going?

Ans: The world is moving towards low fertility rate and greater life expectancy.

1.2 Model life expectancy (40pt)

1. (2pt) Display the distribution of life expectancy. How does it look like? Does it suggest you should use log-transformation? Explain!



The distribution of life expectancy is left-skewed. This suggests that we should use log transformation to transform skewed data to approximately conform to normality.

Reference: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4120293/#:~:text=The%20log%20transformatio

2. (4pt) Create a model where you explain life expectancy with just time

life expectancyt = B0 + B1 · t

where t is time (year). Use year-2000 instead of just year for time.

```
##
## Call:
## lm(formula = lifeExpectancy ~ year, data = updated_gmdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.350  -7.603   2.505   8.042  18.542
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 67.358008   0.109226  616.68   <2e-16 ***
## year         0.308758   0.005441   56.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.14 on 11616 degrees of freedom
## Multiple R-squared:  0.217,  Adjusted R-squared:  0.217
## F-statistic:  3220 on 1 and 11616 DF,  p-value: < 2.2e-16
```

3. (4pt) Why does year-2000 make more sense?

Ans. year2000 makes more sense in order to achieve the line intercept. Since year cannot be zero, finding the value of the intercept is challenging. Thus, we convert the year to display year-2000 so that the predictor variable (year) can take zero value.

4. (4pt) Interpret the results (both B0 and B1).

Ans. B0 is the intercept of the line formulated by the model. The value of inctercept (B0) i.e 67.430274 suggests that when time (year-2000) = 0, the life expectancy value is 67.430274.

B1 is the slopeor the effect of the line formulated by the model. The value of slope (B1) i.e 0.308675 suggests that a positive correlation exists between life expectancy and time where if time increases by 1 unit, life expectancy increases by 0.308675.

5. (4pt) Now let's move to multiple regression: estimate the model where you also add the continent (variable region):

#life expectancyrt = B0 + B1 · t + B1 · regionr

```
## [1] "Americas" "Asia"     "Africa"   "Europe"   "Oceania"
```

```
##
## Call:
## lm(formula = lifeExpectancy ~ year + region, data = updated_gmdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.172  -4.057   0.565   4.041  20.037
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    55.941322   0.123110  454.40   <2e-16 ***
## year            0.304745   0.003574   85.27   <2e-16 ***
## regionAmericas 15.872056   0.182335   87.05   <2e-16 ***
## regionAsia     12.147162   0.169536   71.65   <2e-16 ***
## regionEurope   20.831659   0.180406  115.47   <2e-16 ***
## regionOceania  13.570858   0.264889   51.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.661 on 11612 degrees of freedom
## Multiple R-squared:  0.6624, Adjusted R-squared:  0.6623
## F-statistic:  4557 on 5 and 11612 DF,  p-value: < 2.2e-16
```

6. (4pt) Interpret the results. What do the region dummies mean? What is the reference category? How big is the time trend? Is it statistically significant? Is it different from what you saw in the previous model?

The region dummies indicate that the variable “region” is categorical.

Reference category is Africa

The regression coefficient of year (time trend) is 0.13778. It is statistically significant at 0.1%. The significance of the variable is same however, it’s coefficient has reduced a bit from 0.308 to 0.13 which suggests that it’s effect on lifeExpectancy variable has reduced. lifeExpectancy and Year are still positively correlated.

7. (4pt) As a final result, let’s add two additional variables to the model: log of GDP per capita, and fertility rate. Estimate such a model.

```
##
## Call:
## lm(formula = lifeExpectancy ~ year + region + log(GDP_PC) + fertilityRate,
##     data = updated_gmdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.292  -2.477   0.289   2.724  12.250
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    49.93572    0.50900   98.11   <2e-16 ***
## year            0.13778    0.00355   38.81   <2e-16 ***
## regionAmericas  6.03430    0.15968   37.79   <2e-16 ***
## regionAsia      5.84118    0.15009   38.92   <2e-16 ***
## regionEurope    5.42126    0.20713   26.17   <2e-16 ***
## regionOceania   5.75319    0.22491   25.58   <2e-16 ***
## log(GDP_PC)     2.49027    0.04699   53.00   <2e-16 ***
## fertilityRate  -2.23512    0.04635  -48.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.345 on 8970 degrees of freedom
##   (2640 observations deleted due to missingness)
## Multiple R-squared:  0.8472, Adjusted R-squared:  0.8471
## F-statistic:  7107 on 7 and 8970 DF,  p-value: < 2.2e-16
```

8. (5pt) What do the estimated parameters (betas) for the two new variables tell you?

Ans. The estimated parameters (betas) for the two new variables suggest that:

lifeExpectancy is positively correlated to log(GDP_PC) where one unit increase in log(GDP_PC) causes 2.49027 unit increase in lifeExpectancy

lifeExpectancy is negatively correlated to fertilityRate where one unit increase in fertilityRate causes 2.23512 unit decrease in lifeExpectancy.

Both new variables are statistically significant and have made the model better as the R-square value has increased.

9. (5pt) If you did it correctly, you noticed that Europe was the leading region in Question 5. But now Americas is leading the pack in terms of the value of the region dummy–the dummy for Europe is only 4th largest. Explain why adding additional variables made the ranking of continents to look different.

Adding additional variables made the ranking of continents to look different because both the variables are statistically significant at 1% and had an effect on the lifeExpectancy.

10. (4pt) Based on all the models you have done so far: which continent has the highest life expectancy? Which one the lowest?

###Note: you are welcome to check the group averages too if you wish, but the question asks about the models. See we expect some argumentation based on the models.

Ans. Based on the last model, America had the highest life expectancy.I am basing my answer on the last model as it was the best, which we can tell by looking at the R square.

Africa has the lowest lifeExpectancy since it is the reference category.

2 Find Cheap Restaurants (50pt)

1. (5pt) Load the data and perform basic sanity checks. Ensure you know the variables. Check for missings and unreasonable values and clean the data as necessary.

```
## [1] 168   6
```

```
## [1] "Restaurant" "Food"       "Decor"      "Service"    "East"
## [6] "Cheap"
```

```
## [1] FALSE
```

```
## [1] FALSE
```

```
## [1] FALSE
```

```
## [1] FALSE
```

There are no missing and unreasonable values in the dataset.,

2. (5pt) Your task is to predict if a restaurant is cheap or not. Which type of model, linear or logistic regression do you think is suitable for this task? Explain!

Logistic Regression is a more appropriate model for this task since this is classification problem where the prediction can take up only 2 values.

3. 3. (20pt) Now build the model. Include all the variables you consider relevant for this task. Estimate the model and interpret the statistically significant results. Do your results align with common sense?Ensure you interpret the right type of effects.

```
##
## Call:
## glm(formula = Cheap ~ Food + Decor + Service + East, family = binomial(),
##     data = rest)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1314  -0.6407  -0.2163   0.6142   2.3426
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 19.44989    3.19810   6.082 1.19e-09 ***
## Food        -0.47434    0.20283  -2.339   0.0194 *
## Decor       -0.63285    0.14141  -4.475 7.63e-06 ***
## Service      0.04739    0.19262   0.246   0.8057
## East        -0.33058    0.43586  -0.758   0.4482
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.12  on 167  degrees of freedom
## Residual deviance: 139.70  on 163  degrees of freedom
## AIC: 149.7
##
## Number of Fisher Scoring iterations: 5
```

According to the above summary, we can make the following interpretations:

Statistical Significance:

1. The variable 'Food' is statistically significant at 5%.

2. The variable 'Decor' is statistically significant at 0.1%.

The variable 'Service' is not statistically significant.

The variable 'East' is also not statistically significant.

Correlation:

1. The variable 'Cheap' is negatively correlated to 'Food' with. This means that as food rating increases, the restaurant becomes more expensive.

2. The variable 'Cheap' is negatively correlated to 'Decor'. This means that as decor rating increases, the restaurant becomes more expensive.

3. The variable 'Cheap' is positively correlated to 'East'. This means that as Service rating increases, the restaurant becomes cheaper.

While the correlation between variables 'Cheap' and 'Food' and 'Cheap' and decor make sense, the correlation between 'Cheap' and 'Service' does not make sense as better the service, more expensive the restaurant should be.We cannot comment on the correaltion between 'Cheap' and 'East' as I cannot comment

4. (20pt) You are going out with a few friends and feeling hungry, and would like to have lunch at a not-too-expensive Italian place. You find there are two new places with the following scores and locations:

```
##            Restaurant Food Decor Service East      pred
## 1 Assagio Ristorante   23    17      22    0 -1.175805
## 2             Altura   18    15      24    1  2.225772
```

```
##            Restaurant Food Decor Service East pred
## 1 Assagio Ristorante   23    17      22    0    1
## 2             Altura   18    15      24    1    0
```

What does your model predict–is any of these two restaurants a cheap place? Use the model you made above to find it out.

According to my model, Assagio Ristorante is a cheaper place. Altura on the other hand, is not a cheap restaurant.

3 Theoretical questions (20pt)

1. (6pt) Describe one real-life applications in which logistic regression may be useful, one in which linear regression is useful, and one in which prediction is useful. Describe the response, as well as the predictors. Explain your answer.

Real-life application of Logistic Regression: Predicting whether it will rain today or not based on various parameters

Real-life application of Linear Regression: Predicting the price of a houses based on factors such as sqft, number of bedrooms, location, etc.

Real-life application of prediction: Detecting sickness in healthcare; Predicting House Value;

When two suspect two variables to be causally correlated, where one variable affects the other, the affecting variable is called 'predictor' and the affected variable is called 'response'.

Reference: (OpenIntro Statistics Fourth Edition, David Diez)

2. (5pt) Think about analyzing regression results. What does this mean: A coefficient is statistically significant at 5% confidence level?

When a coefficient is statistically significant at 5% confidence level, it indicates strong evidence against null hypothesis that there is less that there is less than 5% probability of the results being random and by chance.

Reference: https://hbr.org/2016/02/a-refresher-on-statistical-significance, https://www.simplypsychology.org/p-value.html#:~:text=A%20p%2Dvalue%20less%20than,and%2

3. (9pt) You are network security manager. Your network has recently suffered from various attacks and intrusions and now you are evaluating to introduce a new login method, either method L1 or method L2. The login will distinguish between approved users (A) and intruders (I) based on passwords, biometrics and other data. The small-scale evaluation you did produced the following results:

(a) (4pt) Show the confusion matrices for methods L1 and L2. Do it as markdown tables.

Confusion Matrix for L1:

	Actual Login (A)	Actual Login (I)
Predicted (A)	3	3
Predicted (I)	0	4

Confusion Matrix for L2:

	Actual Login (A)	Actual Login (I)
Predicted (A)	2	0
Predicted (I)	1	7

(b) (5pt) Compute accuracy, precision, recall for both models.

To stay on the same page, let's take "I" as positive.

For L1 Model:

Accuracy = ((TP + TN)/(TP + TN + FP + FN)) * 100 = ((4 + 3)/(4 + 3 + 3 + 0)) * 100 = 70

Accuracy of L1 model is 70%

Precision = (TP)/(TP + FP) = (4)/(4 + 0) = 1

Precision of L1 model is 1

Recall = (TP)/(TP + FN) = (4)/(4 + 3) = 0.57

Recall of L2 model is 0.57

For L2 Model:

Accuracy = ((TP + TN)/(TP + TN + FP + FN)) * 100 = ((7 + 2)/(7 + 2 + 1 + 0)) * 100 = 90

Accuracy of L2 model is 90%

Precision = (TP)/(TP + FP) = (7)/(7 + 1) = 0.875

Precision of L2 model is 1

Recall = (TP)/(TP + FN) = (7)/(7 + 0) = 1

(c) (6pt) Which login method, L1 or L2 will you recommend the management to implement?

Explain your reasoning.

When talking about security, it is most crucial to ensure that intruders are not given access.From the confusion matrix, I can tell that method L2 correctly identified all intruders as intruders while method L1 identified 3 intruders as approved, which is not at all good from the security standpoint. Thus model L2 is better. Additionally, the accuracy of model L2 is much better than that of L1.