



Mock Final Examinations

Date / Start Time	Mock Paper
Course	ECON145 Introductory Data Analytics in Healthcare
Groups	G1

INSTRUCTIONS TO CANDIDATES

- 1 The time allowed for this examination paper is **TWO (2) hours**.
- 2 This is a **CLOSED BOOK** test. **No** computers, tablets, smartphones, smart watches, graphical, or programmable calculators are allowed in this examination.
- 3 This examination paper comprises **TWENTY (20)** pages, including this instruction sheet and **TWO (2)** blank pages.
- 4 You are required to answer **ALL** questions.
- 5 You may use a calculator for performing calculations.
- 6 You are required to return the **full set of question papers** at the end of the examination, together with **any extra papers** you requested during the examination.

Campus ID: _____ Seat No: _____

	Marks	Awarded
Section I	32	
Section II	78	
TOTAL	100	
	Grand Total	

Section I: Multiple Choice Questions

Choose 1 (one) best-suited answer for each question. Fill your answers in the following table. Markings on options in the questions will NOT be graded. Each question carries two marks.

Question	Answer
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	

**** Only 16 mock questions are provided. Actual paper contains 20 MCQ questions.**

Q1. Clinic A offers 5 key services and has 8 medical doctors. Each doctor can provide service in any of the key services. At the end of the year, the clinic wants to visualize the total revenue contributed by each medical doctor, along with his/her individual proportions of contribution in each service category. Which of the following is the **most appropriate** visualization for this purpose?

- A. Pie Chart
- B. Line Chart
- C. Column Chart
- D. Stacked Bar Chart

Q2. For each surgery in hospital A, the patient ID and ASA_score of the patient are captured. The Patient ID is an 8-digit unique identifier of the patient which does not indicate any sequence based on the digits. The ASA_score is an assessment of the patient's overall health based on five classes (level I to V), where level I indicates a completely healthy patient. Which of the following best describes the type of data when analysing the data?

	Patient ID	ASA_Score
a.	Numeric (ratio)	Categorical (ordinal)
b.	Numeric (ratio)	Categorical (nominal)
c.	Categorical (nominal)	Categorical (ordinal)
d.	Categorical (ordinal)	Categorical (nominal)

Q3. The following are statements about applying analytics to healthcare in the real-world context. How many of the statements are true?

- i. Only the structured data (e.g., data in databases or well-formed files) can provide value as it is difficult to analyse unstructured data (e.g., text, images data).
- ii. Sensors' data can be used to provide real-time monitoring patients who are recuperating at home, e.g., a home-recovery programme for COVID-19
- iii. A deployable analytics project should include every stage of data analytics (i.e., descriptive, diagnostic, predictive, prescriptive, and embedded).
- iv. The value of analytics lies in collecting as much data as possible.

- A. 1
- B. 2
- C. 3
- D. 4

Q4. How many of the following statements about Exploratory Data Analysis is true?

- i. It is important as it gives us a definite direction to our analysis.
- ii. It allows us to preliminarily assess the appropriate model to use.
- iii. It can be skipped if we are dealing with big data since the volume is too large.
- iv. It allows us to understand if the data is sufficient or appropriate to answer the business questions.

- A. 1
- B. 2
- C. 3
- D. 4

Q5. Which of the following data are examples of structured data?

- i. Name of patient stored as characters in a database field.
- ii. Name of patient in a referral letter written in Microsoft Word.
- iii. Patient ID stored as numeric field in a database table.
- iv. X-ray image of the patient stored on a computer file system.

- A. i and ii
- B. i and iii
- C. i, iii, iv
- D. i, ii, iii, iv

Q6. For a dataset with five input variables, which of the following are valid similarity measurements for the k-means clustering algorithm?

- i. Length in meters
- ii. Euclidean distance
- iii. Cosine similarity
- iv. Hierarchical clustering

- A. ii only
- B. i and ii
- C. ii and iii
- D. i, ii, iii, and iv

Q7. Given the following k-means clustering analysis code chunk, what is the purpose of `set.seed(1)`?

```
...{r}  
# Running k-means clustering  
set.seed(1)  
kmeans_model <- kmeans(data, center=4, nstart=25)  
...}
```

- A. To ensure that every observation will be assigned to one and only one cluster.
- B. To ensure that every cluster has at least one observation.
- C. To ensure that all runs of the clustering analysis using the same similarity measure and other parameters will produce four clusters.
- D. To ensure that all runs of the clustering analysis using the same similarity measure and other parameters will yield the same cluster assignment.

Q8. Standardization is an important step to be considered before applying the k-means algorithm. Why?

- A. To ensure that all variables have equal importance in similarity comparison.
- B. To get the optimal efficiency in the execution of clustering analysis.
- C. To get the optimal cluster assignment with the minimum within-cluster sum of squares (WSS).
- D. To change the distribution of some variables to handle extreme values such as the outliers.

Q9. The false-positive rate is

- A. the probability that a predicted positive case is actually negative.
- B. the probability that a predicted negative case is actually positive.
- C. the probability that an actual positive case is predicted to be negative.
- D. the probability that an actual negative case is predicted to be positive.

Q10. If we plan to analyze the effectiveness of the National Steps Challenge™ in promoting public health, measured in terms of increased physical activity level, using data collected from the fitness trackers, we must be aware of

- i. outliers due to fraud behaviours
- ii. inaccurate observations due to the sensitivity of the trackers
- iii. selection bias due to the voluntary participation

- A. i and ii
- B. i and iii
- C. ii, and iii
- D. i, ii, and iii

Q11. Which of the following is not a reason that randomized clinical trials (RCTs) are challenging to conduct?

- A. We may not have enough time to wait for the outcomes to be observed.
- B. We can only observe one potential outcome for each subject, i.e., with or without treatment.
- C. We may not convince enough subjects to participate in the RCT.
- D. Sometimes, we may not be able to blind the subjects completely, so the subjects know exactly whether they are in the treatment or control group.

- Q12. Several studies have identified that the incidence of colorectal cancer was associated with (1) older age, (2) having a personal history of inflammatory bowel disease, (3) having a family history of colorectal cancer or colorectal polyps, (4) lack of regular physical activity, (5) a diet high in red meats and processed meats, (6) alcohol consumption, and (7) tobacco use, etc. Based on this information, which are considered risk factors for developing colorectal cancer?
- A. (1) and (3)
 - B. (1), (2), and (3)
 - C. (4), (5), (6), and (7).
 - D. All of them.
- Q13. Which of the following methods do we use to find the best fit line for the data in linear regression?
- A. Minimizing the sum of squared errors.
 - B. Maximizing the likelihood of observing the data.
 - C. Minimizing the Gini index.
 - D. Either A or C.
- Q14. One of the logistic regression models we obtained in the class on Framingham Heart Study has a logit = $-9.0913 + 0.4953 \text{ male} + 0.0679 \text{ age} + 0.0226 \text{ cigsPerDay} + 0.0021 \text{ totChol} + 0.0169 \text{ sysBP} + 0.0067 \text{ glucose}$. Based on this model, being male is associated with
- A. a 0.4953% increase in the probability of developing coronary heart disease (CHD) in the next ten years.
 - B. an $e^{0.4953}$ % increase in the probability of developing CHD in the next ten years.
 - C. an $e^{0.4953}$ increase in the odds of developing CHD in the next ten years.
 - D. an $e^{0.4953}$ times increase in the odds of developing CHD in the next ten years.
- Q15. Which of the following techniques is a supervised learning technique?
- A. Exploratory data analysis
 - B. Propensity score matching
 - C. Classification and regression trees
 - D. K-means clustering
- Q16. In a hospital, an elderly male patient was accidentally given ten times his prescribed dose of insulin when staying in the general ward. Fortunately, there was no adverse outcome after a period of close monitoring, and he was safely discharged as expected for his course of treatment. From this incident, which aspect should the hospital focus on to improve its quality of service?
- A. Access
 - B. Structure
 - C. Process
 - D. Outcome
-

Section II: Contextual Questions

Note: For ALL questions in this section, provide answers in the given SPACE. Answers written elsewhere will NOT be graded. Please round your answer to TWO decimal places for non-integer numbers.

***** Mock paper allocated 78 marks. Actual paper will allocate 60 marks to Section II.**

Q17. [4 marks] Match the scenarios to the **most suitable** big data characteristic. Choose from Velocity, Variability, Veracity, Variety.

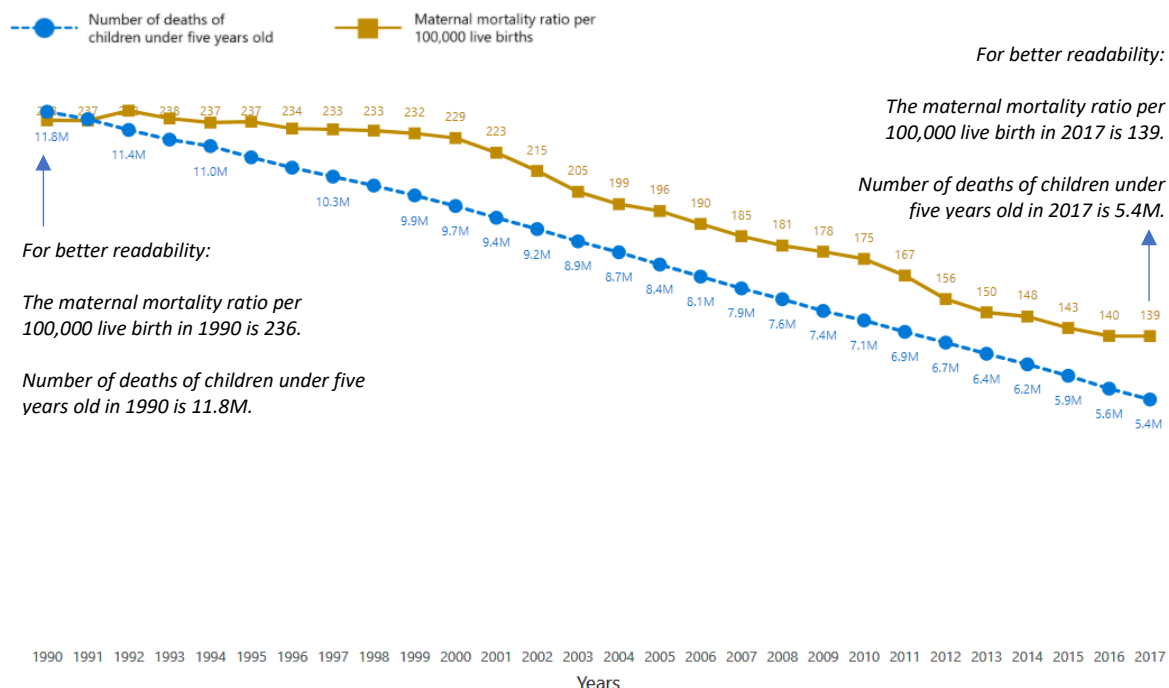
Statement	Characteristic
High number of posts about COVID-19 on social media in a minute.	
Different brands of sensors capturing the room temperatures of different parts of the hospital.	
Airport's data capturing photos, arrival information, body temperatures (high/normal) and COVID-19 declarations of passengers.	
GPS data showing drones delivering medical supplies traveling from one continent to another within a short amount of time.	

Q18. [4 marks] Choose the **most suitable** type of analytics for each of the scenarios. You are to select from this list: *descriptive, diagnostic, predictive, prescriptive, or embedded*.
Note: You may use an option more than once.

Scenario	Type of Analytics
Estimate the bill size of a patient's stay in a hospital based on the type of surgery, choice of services selected and expected length of stay during pre-admission financial counselling.	
Investigate if medical machine of type A has significantly higher breakdown time than machine of type B.	
Monitor and respond to the healthcare conditions of the patients in the wards based on real-time vital signs.	
Evaluate two different polyclinic configurations using simulation for handling patient arrival surges due to a pandemic.	

Q19. Study the following 2 charts about child mortality and maternal mortality.

Note: Maternal mortality ratio: the number of material deaths per 100,00 live births. Maternal deaths as any death of a woman while pregnant or within on year of termination of pregnancy.



Source: IHME

Microsoft, Data Science and Analytics

Figure 19-1: Child and Maternal Mortality over years

(Source: IHME, Microsoft, Data Science and Analytics)

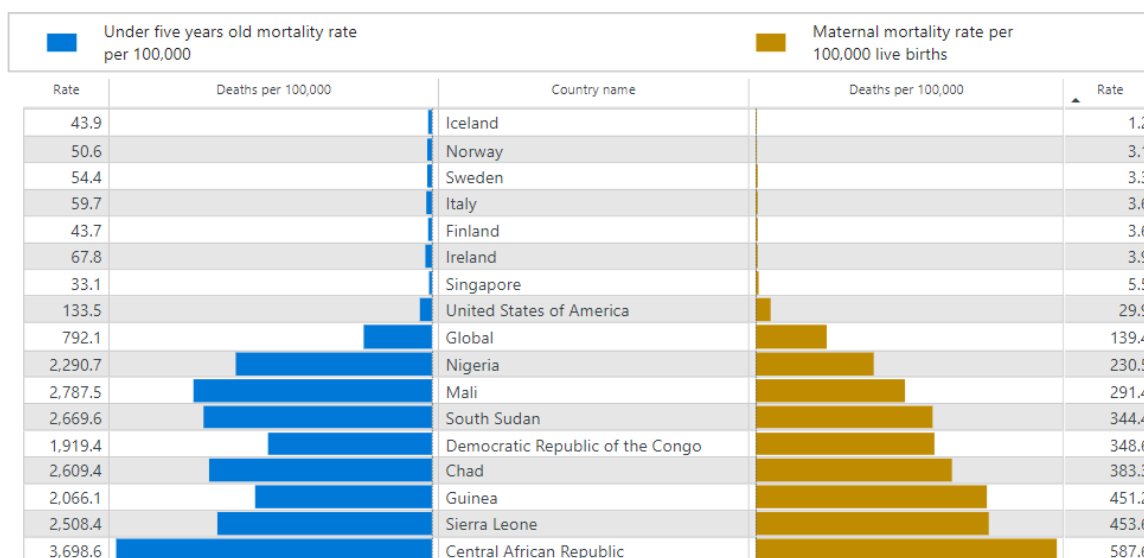


Figure 19-2: Breakdown by country based on data in Year 2017

(Source: IHME, Microsoft, Data Science and Analytics)

- a. **[4 marks]** Based on Figure 19-1, select two statements which are evidently true based on the chart above.
- i. Child mortality and maternal mortality are linearly correlated statistically.
 - ii. Child mortality and maternal mortality are inversely correlated.
 - iii. Child mortality and maternal mortality showed a downwards trend.
 - iv. Child mortality and maternal mortality have reduced by about 54% and 41% respectively from 1990 to 2017.

Answer:	
---------	--

- b. **[4 mark]** Based on Figures 19-1 and 19-2, which of the following is a valid conclusion about child and maternal mortality across the countries?
- i. Reduction in child mortality and maternal mortality are not at the same rates across the countries from 1990 to 2017.
 - ii. Developed countries have statistical significantly lower rates of child and maternal mortality than the underdeveloped countries.
 - iii. Both i and ii.
 - iv. Not possible to conclude.

Answer:	
---------	--

- c. **[4 marks]** In 2017, the average number of death of children under five years old globally is 38,004 but the median is 51,000. State whether you think Child Mortality is right-skewed, left-skewed or balanced. Provide one (1) possible reason for the difference between the two central tendency measures.

Choice	Right-skewed / Left-skewed / Balanced (e.g., Normally distributed)
Possible Reason	

- d. [2 mark] If you like to perform clustering analysis using this data set, which of the following is a valid variable to use? [Select all that applies]
- i. Child mortality
 - ii. Maternal mortality
 - iii. Country
 - iv. Year

Answer:	
---------	--

EXAM-IN-CONFIDENCE

Q20. Read the excerpt of an article given below:



NEW YORK CITY—Wearables company Fitbit is deepening its reach into healthcare with a new premium subscription service for users that offers coaching and personalized insights mined from the health data it collects from 27.3 million users.

Fitbit plans to leverage its consumer health data, collected over 10 years and 27.3 million users, to provide actionable insights on both an individual level and a population level.

The company has collected 228 billion hours of heart rate data from users along with 202 trillion steps, 10.5 billion nights of sleep and 517 billion minutes of exercise. Women who use Fitbit also have logged 42 million periods.

[Only excerpts of the article were selected for the purpose of this examination]

Source: Fierce Healthcare <https://www.fiercehealthcare.com/tech/what-fitbit-s-new-product-and-services-say-about-company-s-health-ambitions> by Heather Landi | Aug 28, 2019

- a. **[6 marks]** Based on the excerpt above, **provide an analytic question** that you can answer using the data collected each of the categories: individual level and population level. For each analytic question, also **describe the data** you will use.

	Your analytic question and data
Individual level	
Population level	

- b. **[6 marks]** Given that heart rate and amount of sleep are associated with health conditions and additional demographic data of the user has been collected. Design a k-means clustering analysis model using age (numeric), weight (numeric), height (numeric), gender (M/F), average heart rate (numeric), amount of sleep in a week (numeric), average number of steps per week (numeric).

Variable(s) requiring (one-hot or dummy) encoding	
Does variables standardization?	<input type="checkbox"/> Yes <input type="checkbox"/> No Reason: _____
Target variable (select one or N/A if not applicable)	
Independent variables (list all applicable ones)	

Q21. Flu epidemics constitute a major public health concern causing respiratory illnesses, hospitalizations, and deaths. The U.S. Centers for Disease Control and Prevention (CDC) and the European Influenza Surveillance Scheme (EISS) detect influenza activity through virologic and clinical data, including Influenza-like Illness (ILI) physician visits. Reporting national and regional data, however, are published with a lag time of one to two weeks. The Google Flu Trends project was initiated to see if faster reporting can be made possible by considering flu-related online search queries, which are available almost immediately.

The CDC publishes the official regional and state-level percentage of patient visits to healthcare providers for ILI purposes every week. Google Trends allows public retrieval of weekly counts for every query searched by users around the world. For each location, the counts are normalized by dividing the count for each query in a particular week by the total number of online search queries submitted in that location during the week. Then, the values are adjusted to be between 0 and 1.

To test the predictability of the Google queries on the spread of flu epidemics, we collected a data set with the following variables.

Variable	Description
<i>Week</i>	The range of dates for the week
<i>ILI</i>	The percentage of ILI-related physician visits for the corresponding week
<i>Queries</i>	The fraction of queries that are ILI-related for the corresponding week, adjusted to be between 0 and 1 (higher values correspond to more ILI-related search queries)

We randomly separated the data set into a training set and a test set. The screenshot of the R Notebook for data loading and preprocessing is provided on the next page.

```

##{r}
flu = read.csv("Data/Flu.csv")
str(flu)
summary(flu)
##

```

```

'data.frame': 469 obs. of 3 variables:
 $ week : chr "2004-01-04 - 2004-01-10" "2004-01-11 - 2004-01-17" "2004-01-18 -
2004-01-24" "2004-01-25 - 2004-01-31" ...
 $ ILI : num 2.42 1.81 1.71 1.54 1.44 ...
 $ Queries: num 0.238 0.22 0.226 0.238 0.224 ...
      week      ILI      Queries
Length:469      Min. :0.534      Min. :0.0412
Class :character 1st Qu.:0.927      1st Qu.:0.1793
Mode :character  Median :1.277      Median :0.2869
                Mean  :1.675      Mean  :0.2997
                3rd Qu.:2.021      3rd Qu.:0.3918
                Max. :7.619      Max. :1.0000

```

```

##{r}
library(caTools)
set.seed(1)
split = sample.split(flu$Queries, splitRatio = 0.80)
split[1:10]
##

```

```

[1] TRUE TRUE TRUE FALSE TRUE FALSE FALSE TRUE TRUE TRUE

```

```

##{r}
fluTrain = subset(flu, split == TRUE)
fluTest = subset(flu, split == FALSE)
str(fluTrain)
summary(fluTrain)
##

```

```

'data.frame': 375 obs. of 3 variables:
 $ week : chr "2004-01-04 - 2004-01-10" "2004-01-11 - 2004-01-17" "2004-01-18 -
2004-01-24" "2004-02-01 - 2004-02-07" ...
 $ ILI : num 2.42 1.81 1.71 1.44 1.04 ...
 $ Queries: num 0.238 0.22 0.226 0.224 0.216 ...
      week      ILI      Queries
Length:375      Min. :0.534      Min. :0.0412
Class :character 1st Qu.:0.914      1st Qu.:0.1773
Mode :character  Median :1.281      Median :0.2842
                Mean  :1.673      Mean  :0.2982
                3rd Qu.:2.020      3rd Qu.:0.3891
                Max. :7.619      Max. :1.0000

```

```

##{r}
str(fluTest)
summary(fluTest)
##

```

```

'data.frame': 94 obs. of 3 variables:
 $ week : chr "2004-01-25 - 2004-01-31" "2004-02-08 - 2004-02-14" "2004-02-15 -
2004-02-21" "2004-05-02 - 2004-05-08" ...
 $ ILI : num 1.542 1.324 1.307 0.757 0.804 ...
 $ Queries: num 0.2377 0.2072 0.2417 0.0611 0.0491 ...
      week      ILI      Queries
Length:94      Min. :0.586      Min. :0.0465
Class :character 1st Qu.:1.015      1st Qu.:0.1879
Mode :character  Median :1.250      Median :0.3001
                Mean  :1.684      Mean  :0.3056
                3rd Qu.:2.017      3rd Qu.:0.4031
                Max. :5.423      Max. :0.7596

```

- a. **[4 Marks]** The ratio used to split the training and test data sets is _____. The maximum value of *ILI* in the training set is _____.

We first built a simple linear regression model and obtained the following results from R.

```

{r}
fluLR = lm(ILI ~ Queries, data = fluTrain)
summary(fluLR)

```

```

Call:
lm(formula = ILI ~ Queries, data = fluTrain)

Residuals:
    Min       1Q   Median       3Q      Max
-1.5428 -0.3948 -0.0542  0.3236  2.5586

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0217    0.0721     0.3    0.76
Queries      5.5386    0.2127    26.0   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.665 on 373 degrees of freedom
Multiple R-squared:  0.645,    Adjusted R-squared:  0.644
F-statistic: 678 on 1 and 373 DF,  p-value: <2e-16

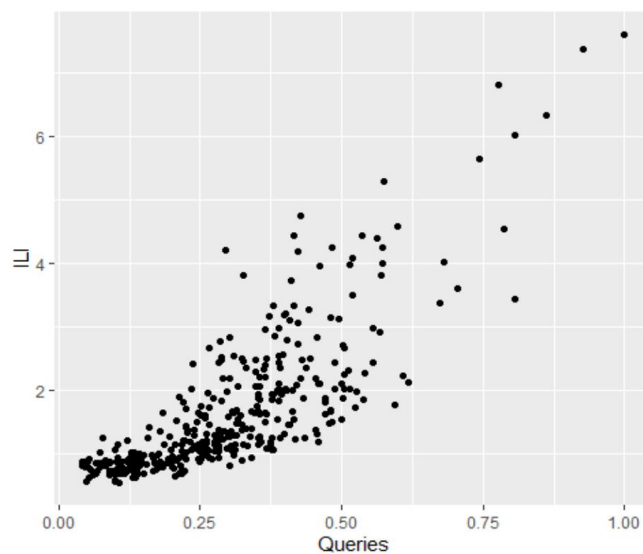
```

- b. **[4 Marks]** The R-squared of the model is _____. The estimated coefficient of *Queries* is _____.
- c. **[4 Marks]** Answer true (T) or false (F) for the following statements.

The model's R-squared is too low for the model to be useful.	
We can conclude that larger values of <i>Queries</i> lead to more <i>ILI</i> physician visits.	
The variable, <i>Queries</i> , is significant at the 0.05 level.	
The variable, <i>Queries</i> , is significant at the 0.001 level.	

- d. **[4 Marks]** For the second data point in the test set, the above linear regression model would predict _____% of *ILI*-related physician visits. For the same week, the baseline model would predict _____% of *ILI*-related physician visits.

We could see that the simple linear regression model is not very accurate. To improve the model, we made the following scatter plot to check the relationship between *ILI* and *Queries*.



- e. **[6 Marks]** From the plot, we observed a clear positive relationship between the two variables. However, the relationship is not quite linear. What can we do to improve the model's accuracy? For each of the following suggested options, answer true (T) for correct ones and false (F) for incorrect ones.

Apply a logarithmic transformation on <i>ILI</i> .	
Apply a logarithmic transformation on <i>Queries</i> .	
Apply a square-root transformation on <i>ILI</i> .	
Apply a square-root transformation on <i>Queries</i> .	
Include the square of <i>ILI</i> as another independent variable.	
Include the square of <i>Queries</i> as an additional independent variable.	

Instead of using linear regression, we could also train a regression tree model for this problem. We conducted 10-fold cross-validation to determine the optimal parameter for the tree model. The corresponding R code and output are provided below.


```

{r}
fitControlFHS = trainControl(method = "cv", number = 10)
cpGrid = expand.grid(.cp = (1:20)*0.001)
set.seed(1)
train(IL1 ~ Queries, data = fluTrain, method = "rpart", trControl = fitControlFHS,
tuneGrid = cpGrid)

```

CART

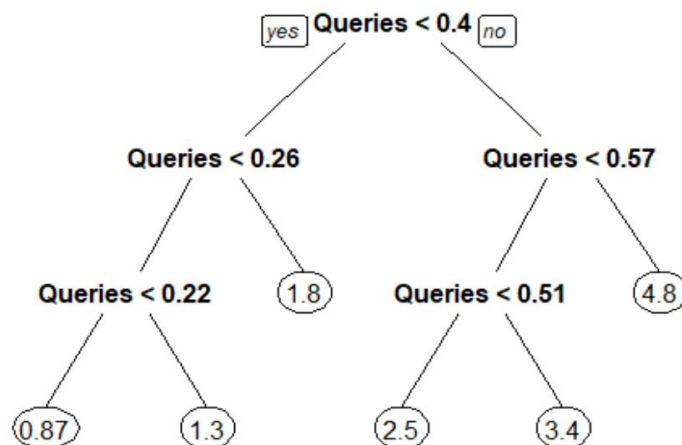
375 samples
1 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 339, 339, 339, 335, 337, 337, ...
Resampling results across tuning parameters:

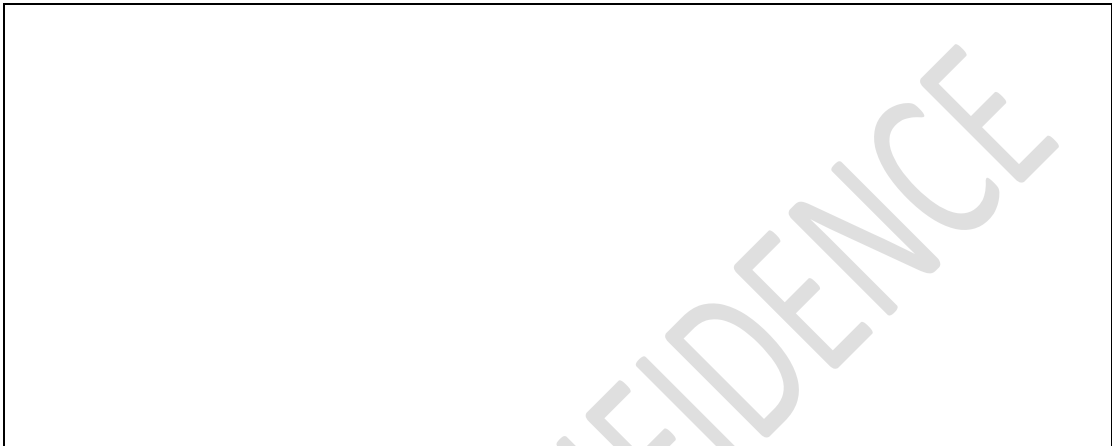
cp	RMSE	Rsquared	MAE
0.001	0.7089	0.5854	0.4801
0.002	0.7032	0.5890	0.4747
0.003	0.7027	0.5895	0.4714
0.004	0.7046	0.5891	0.4732
0.005	0.6993	0.5966	0.4725
0.006	0.6947	0.5983	0.4701
0.007	0.6959	0.5958	0.4698
0.008	0.6941	0.5978	0.4709
0.009	0.6941	0.5978	0.4709
0.010	0.6951	0.5969	0.4713
0.011	0.6965	0.5955	0.4745
0.012	0.6968	0.5951	0.4766
0.013	0.6990	0.5935	0.4801
0.014	0.7029	0.5907	0.4832
0.015	0.7029	0.5907	0.4832
0.016	0.7078	0.5818	0.4896
0.017	0.7106	0.5778	0.4924
0.018	0.7106	0.5778	0.4924
0.019	0.7106	0.5778	0.4924
0.020	0.7106	0.5778	0.4924

- f. [2 Mark] The optimal complexity parameter based on the above cross-validation should be _____ (do **NOT** round this answer).

Using the optimal parameter found above, we trained a regression tree from the training data set. The final tree is plotted below.



- g. **[4 Marks]** There are _____ splits and _____ leaf nodes in this tree model.
- h. **[2 Mark]** For the fifth data point in the test set, the above tree model would predict _____ % of ILI-related physician visits.
- i. **[6 Marks]** Comparing this regression tree model with the linear regression model (in part b), which one do you prefer and why?



EXAM-IN-CONFIDENCE

- This page is blank -

EXAM-IN-CONFIDENCE

----- **End of Paper** -----