CS381/780 Data Analytic Mid-term Exam 3/17/2021

Instruction: For multiple choice questions, clearly circle one of the choice; for all other questions, write your answer right below the questions. All questions carry the same weights.

Name: Seth Marcus

I lost the images while converting it to Microsoft word in order to take the exam.

Question 1: Given the following database tables, write a SQL statement to list the name of employees in the Finance department who was hired before year 2020

Answer:

```
Select E.emp_name
From employees as E join department as D
  On E.dep_id = D.dep_id
Where Extract(YEAR FROM E.hire_date) < 2020 AND upper(D.dep_name) =
upper('Finance');
```

Note: I am using Postgressql. Why? That is what I am currently using outside of class. All it says is SQL, does not mention a particular type.

Question 2: Based on the database tables in previous question, write a SQL statement to list all employees of grade 3 and 4 whose department location is in Sydney

Answer:

```
Select e.emp_name , e.salary , d.dep_name , d.dep_location, sg.grade
From employees AS E join department as D
  ON E.dep_id = D.dep_id and upper(d.dep_location) = upper('Sydney')
  Join salary_grade as sg
  On E.salary >= sg.min_sal and E.salary <= sg.max_sal
Where sg.grade >= 3;
```

Note: I am using Postgressql. Why? That is what I am currently using outside of class. All it says is SQL, does not mention a particular type. I chose postgres.

Question 3: What does EDA stands for and what are some of the typical tasks in EDA?

Answer: **E**xploratory **D**ata **A**nalysis. Typical tasks include but are not limited to: spotting errors in the data, mapping out the underlying structure of the data, estimate parameters and figuring out the associated confidence intervals or margins of error, establish a mode that can be used to explain the data with minimal predictor variables, test a hypotheses/check assumptions to the related model and listing anomalies and outliers amongst others.

Question 4: What is the purpose of building a Data warehouse? List out 3 differences between a Data Warehouse and a traditional database?
Answer:
Purpose of DW: Data in DW is cleansed data used for reporting and analysis. Special data management facility whose purpose is to create reports and analysis to support managerial decision making. DW also enable a consolidated view of corporate data, all cleaned and organized. They are also designed to make reporting and querying simple and efficient (Keep It Simple Stupid). The sources of data are operational systems, and external data sources. The data warehouse needs to be updated oftenlly with new data to keep it useful (garbage in, garbage out). Data from a data warehouse provides useful input for data mining activities.

Data Warehouse	Traditional Database
Lower granularity	Highly granular data (i.e all activity and transaction details)
grow from the database itself and the data is rolled up and appended every so often (no set rules, depends on circumstance).	The size of the dataset grows with activity and transactions (even a delete increases the size since it is a soft delete and the transaction is recorded in the logs).
Complexity: typically organized around a large fact table, and many lookup tables.	Highly complex with dozens or hundreds of data files, linked through common data fields.

Question 5: On comparing a typical database in a OLTP system versus a data warehouse in a OLAP system, which of the following are true?

1. Data warehouse has more granularity on the transaction when compared with a traditional database because it needs more details for business decision support.
2. Updates on a OLTP system typically happens live while updates on OLAP system can be on a daily, weekly or monthly basis.
3. Data Mart of an enterprise refers to the collection of different Data Warehouse from different department combined together just like Wal-Mart inventory is coming from their warehouses located on different regions of the country

A. Only 1
B. Only 2
C. 1 and 2
D. 2 and 3
E. 1, 2 and 3

Answer: E. 1,2, and 3

Question 6: Who will be responsible for running the ETL process and making sure the results make sense?

A. Data Engineer
B. Data Analyst
C. Data Scientist

Answer: A. Data Engineer

Question 7: On given the following weights (lbs) dataset: {80, 100, 100, 80, 110, 70, 90}.

Answer the following questions (show your calculation)

A. What is the mean?
   a. 90
B. What is median?
   a. 90
C. What is the mode?
   a. 80
D. What is the standard deviation?
   a. 13.09307
   b. For D, I assumed this is the population. The sample deviation is 14.14214

Question 8: What is selection bias and how can you avoid it?

Answer:

Selection bias is a distortion in a measure of association due to a sample selection that does not accurately reflect the target population. Selection bias can occur when investigators use improper procedures for selecting a sample population, but it can also occur as a result of factors that influence continued participation of subjects in a study. In either case, the final study population is not representative of the target population – the overall population for which the measure of effect is being calculated and from which study members are selected.

Question 9: Which of the following statements is/are true about Type-1 and Type-2 errors?

A. Only 1
B. Only 2
C. 1 and 2
D. 2 and 3
E. None of the above

Answer: A. Only 1

Question 10: Which of the following are true

1. In a negatively skewed distribution, the mean will be less than the median
2. In a positively skewed distribution, the median is larger than the mean
3. In a normal distribution, the mean and the mode are the same

A. Only 1
B. Only 2
C. Only 3
D. 1 and 3
E. 2 and 3

Answer: C. Only 3. See this for why the standard "rule of thumb" is false. (namely why 1 isn't always true).

Question 11: In the following double hump distribution, which of the following are true

A. 1 and 2
B. 1 and 3
C. 2 and 3
D. All of the above

Answer: B. 1 and 3

Question 12: Select which of the following are true

1. The net worth distribution of a retired community in California will have a higher mean than the distribution for the whole country
2. The net worth distribution of a retired community in California will have a higher standard deviation than the distribution for the whole country
3. The net worth distribution of a retired community in California will have a higher kurtosis than the distribution for the whole country

A. Only 1
B. Only 2
C. 1 and 2
D. 1 and 3
E. None of the above

Answer: D. 1 and 3

Question 13: Suppose you are given the following plots 1-4 (from left to right) and you want to compare their Pearson correlation coefficients. Which of the following are true (including the sign)?

A. Only 1
B. Only 2
C. 1 and 2
D. 2 and 3
E. None of the above

Answer: A. Only 1

Question 14: Name and describe 3 common sampling methods and their corresponding pros and cons.

Answer:

A. Random Sampling
    a. Select a sample of data at random, not convivence, but rather truly random and according to the laws of probability if large enough, will sufficiently represent the population.
        i. Pros:
            1. Can do more advanced statistical techniques to delve deeper into the answers that the data can provide.
        ii. Cons:
            1. Need a large dataset.
            2. Very hard to actually get a truly random sample.
B. Convenience Sampling
    a. Select people that are really easy to find/get answers from.
        i. Pros:
            1. Cheap and easy
        ii. Cons:
            1. Likely to not be representative of the population despite the sample size.
C. Systematic Sampling
    a. Way to sample a set of data done arbitrarily for each case.
        i. Pros:
            1. More straightforward than Random sampling
            2. Can create a model to sample any set of data you want.
        ii. Cons:
            1. May introduce bias since the sampling methods are made arbitrarily for each case.

Question 15: In regards to filling missing values, which of the following are true?

1. Removing the problematic rows may introduce systematic bias.
2. Using mean to fill in the missing values ia always preferable because it captures on average the most common possibility.
3. Using median is always preferable than the mean when there are many outliers in the data set because it is less sensitive.

A. Only 1
B. 1 and 2
C. 1 and 3
D. 2 and 3
E. All are correct

Answer: A. Only 1. “Always”; it is never “Always” with any of these decision questions. It really depends case by case.

Question 16: Explain what is Bias and Variance Trade-off.

Answer:

Bias vs. Variance:

Bias means your model is intrinsically wrong (off, biased, incorrect, etc.) and that you will not fit the data well. If you use a too simplistic model, you will have high bias. On the other side of the coin, using a more complicated model, you will have low bias. However, your model will not generalize will to testing dataset(s) (out of sample data). The ‘variance’ of your prediction will be high. What we need to do is find the right (or good enough) balance without overfitting or underfitting the data.

Question 17: Adding additional variables to a linear regression model will result in

A. Only 1 is correct
B. Only 2 is correct
C. Only 3 is correct
D. None of the above

Answer: C. Only 3 is correct.

Question 18: Suppose you are given a dataset, which of the followings are considered as good common practice in building model.

1. Use a big portion of the data points from your dataset to fit your model, and reserve the rest for testing your model.
2. Take different samples from your dataset to build different versions of a model may not be a good idea because you will not be sure which one of these models will be right.
3. Use as many features as possible from your dataset because different features may be able to explain the different part of the behavior of the target variables.

A. Only 1
B. Only 2
C. 1 and 2
D. 1 and 3
E. 1, 2 and 3

Answer: A. Only 1.

3 is false because we only want to use relevant features (I love ice cream should have no impact on how good of a basketball player I am, e.g. Should not be used in a model for basketball players).

Question 19: In hypothesis testing, which of the following are true

A. Only 1
B. Only 2
C. Only 3
D. 1 and 2
E. 2 and 3

Answer: A. Only 1

2 is false. Regarding 3, we never accept the alternative hypothesis; we merely reject the null hypothesis.

Question 20: In regards to Linear and Logistic Regression, which of the followings are true?

1. When we are presented 10 different models with different number and choices of independent variables in Linear Regression, we always want to pick the model that has highest R-square as R-square determines how good the fit is.
2. The range of an odd of an event is between 0 and 1 because by definition probability has to be between 0 and 1.
3. R-square is used as a model performance metrics for both Linear and Logistic Regression.

A. Only 1
B. Only 2
C. 1 and 2
D. 1 and 3
E. None of the above

Answer: C. 1 and 2