**DS 620 – Week 5 Assignment**

**Ritesh Kumar B**

1 . Performance Evaluation. Kiwi Analytics is assessing two different training programs for its consulting employees. One group of 50 employees used training method A for varying numbers of hours and another group of 50 employees used training method B for varying numbers of hours. Then these employees were evaluated based on their performance in job-related tasks. The resulting data from this experiment is displayed in the following two scatter charts.

Chart, scatter chart

Description automatically generated

Which of the following are accurate statements based on these data? Select all that apply.

1. Training method A and training method B are equally effective at improving evaluation scores.
2. Training method A is more effective than training method B because the observations in the first scatter chart exhibit less variability around the linear trendline than the observations in the second scatter chart.
3. Correlation does not imply causation and therefore there is no meaningful conclusion from these charts.
4. The strength of the linear relationship between hours of training and evaluation score is the same for both training method A and training method B.

ANSWER:

Training method A is more effective than training method B because the observations in the first scatter chart exhibit less variability around the linear trendline than the observations in the second scatter chart.

-----------------------------------------------------------------------------------------------------------------------------------------

2. Types of Missing Data. Label each of the following descriptions based on the type of missing data being described: missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR).

a. A survey is conducted that asks a sample of people their age, highest level of education completed, and preference of political party. The results indicate that many entries for highest level of education completed are missing, and that respondents were more likely to omit this response if they had a high school education or less. **(MNAR)**

b. An economic study is conducted asking people to report their profession, job title, and city of residence. Approximately 10% of the responses for job title are missing, and it is observed that men are less likely to provide their job title than women. **(MAR)**

c. Students at a local college are given a survey to collect data on their study habits and preferred learning styles. Students complete the survey by filling out a Scantron form on which they fill in circles to specific multiple-choice questions. Due to Scantron-reading errors, five % of the question responses are not able to be read correctly and so are recorded as missing. **(MCAR)**

d. A clinical study is being done to judge the efficacy of a new drug in reducing a patient’s high blood pressure. Each month study participants must each undergo a stress test where they run on a treadmill at maximum effort for 90 seconds, after which the patient’s blood pressure is recorded. However, each patient is screened prior to the stress test and if their resting blood pressure is too high, they do not complete the stress test and the blood pressure reading for that patient is missing. **(MNAR)**

-------------------------------------------------------------------------------------------------------------

3. Tax Data by County. The file TaxData contains information from federal tax returns filed in 2007 for all counties in the United States (3,142 counties in total). Create an Excel Table based on these data. Using Excel Table functionality, answer the following questions.

a. Which County had the largest total adjusted gross income in the state of Texas?

b. Which county had the largest average adjusted gross income in the state of Texas?

Answer is submitted in the form of Excel Sheet

----------------------------------------------------------------------------------------------------------------------------------------

4. More Univariate Analysis of EJB Data. In the EJB example discussed within the chapter, univariate analysis was demonstrated on some of the variables. We continue this analysis in this problem.

a. Using a PivotChart, construct the relative frequency distribution of records over the values of the Category variable. Describe your findings.

b. Using a PivotChart, construct the relative frequency distribution of records over the values of the New Customer variable. In the PivotTable, relabel a “No” value for New Customer as “Existing” and a “Yes” value as “New.”

Describe your findings:

* Orange was the most sold Flavor amongst all others, while Tomato was the least sold Flavor followed by Beet varying with a very small difference.
* Juices are the most ordered category in the sales.
* About 17.5% of the customers rated a “5” for the Product Satisfaction Rating, while only about 3.9% of them rated “1”
* NE is the distribution center that is bringing in most of the orders contributing to 15.68% of the overall, followed by ND and NM & ID.
* MS is the distribution center with the least number of orders when compared side by side with other centers
* Most (21.7%) of the customers gave a high rating of “5”, and only about 10% of the rating the least for the Service Satisfaction part.
* About 79% of the orders the business gets is from the current/existing customers, than compared to new customers which only contributes to about roughly 21%
* All in all, the service and product satisfaction rating are overall on the higher end, which is a good sign for the business.
* While the frequency of orders from new customers is very low and something to be worked upon

-----------------------------------------------------------------------------------------------------------------------------------------

5. Smartphone Sales. The file Smartphone contains data on the monthly sales revenue for a smartphone manufacturer.

a. Create a line chart to depict the sales time series at the annual level.

b. Create a line chart to depict the sales time series at the quarterly level.

c. Create a line chart to depict the sales time series at the monthly level.

d. What insight do each of these three views provide?

* Sales by just year shows us that the business in upward trend, however but only with a minute increase
* Sales by quarter, gives us some better insights into sales as compared to the annual chart.
* We can say that every 2nd qtr., the business goes down and does not perform well. This pattern is observed for all the years till date. While it picks back up during the 3rd qtr., and q4 being the best qtr., for all the 4 consecutive years.
* Moving to the detailed chart for years, qtrs. And months, it gives us a more detailed view as to how the business performs for every given period of the year.
* Here, we saw that q4 brings the highest business, but here we can also observe that Q4 is the quarter with most business fluctuations.
* The sales seem to be inconsistent for every month/qtr and year.
* However, we can observe that Nov is the month that hit the minimum sales for the year 2018,2019 & 2020, while it was October for 2021.
* Maximum Sales for any given year are the most in Q3 & Q4.

----------------------------------------------------------------------------------------------------------------------------------------