**Module 7: Critical Thinking – Final Research Paper**

Balajee Sivakumar

Data Analytics, Colorado State University (CSU) Global

MIS581-1: Capstone - Business Intelligence and Data Analytics

Dr. Osama Morad

10/01/2023

**Abstract**

This capstone project analyzes my selected dataset against the hypothesis that I have set during the beginning of this course and then provides the outcome of the same. For my Capstone project, I will be using the historical and current data to analyze different real estate data for a particular state – Connecticut so that our construction organization can benefit from this analysis. I have used references from different reliable sources like google scholar, CSU global library, etc., to gather the information required for this project. Our construction company’s first goal is using the provided data to understand the change in the average “Sales Ratio” over a 20-year period. The second goal is to run prediction models on this data for forecasting the future home prices which will help them come up with their pricing roadmap. These goals will be achieved by using the research design methods and methodologies mentioned in this document.

*Keywords:* Analytics, Capstone Project, Hypothesis, Real estate Sales analytics, Privacy

**Final Research Paper**

As part of this week’s critical thinking assignment, I will be providing the Final research paper. Firstly, I will be mentioning the proposal explaining the reason behind my research along with the problem statement. Secondly, I will be providing an explanation about my selected dataset and the reason behind my choice. Thirdly, I will be providing an overview of the hypotheses. Fourthly, I will be providing a complete literature review of my selected scholarly articles that are supporting my research. Fifthly, I would be describing the research method (models and tools) used to analyze the selected dataset to test my hypothesis. Sixthly, I would be addressing the security, privacy, and ethical concerns while handling the data. Finally, I will be explaining the result of my hypothesis along with the factors may have influenced the same. As mentioned by John Adams, et al., (2014), It is essential for any project to understand what has already been done in the specific topic that we have chosen also covering the wider subject area of that topic as well. This help us as a researcher to gain clear understanding of the need and importance of our research and become confident about our research.

**Proposal and Problem statement**

As mentioned by McGaghie, et. al., (2001), problem statement is an essential element that conveys the issues and context that give rise to a study. The organization that we have selected is a construction company based out of Connecticut. They are currently running into losses post the covid pandemic because of multiple reasons like higher raw materials cost, delay in the supply chain, pricing strategy not being in line with the market standards, etc., As per Geisler, J. (2020, May 1), in critical times like the COVID-19 pandemic situation, the followers want their leader to be realistic and optimistic. But our construction company leadership team had to increase their sales pricing temporarily without doing proper data analysis in order to compensate for the above-mentioned factors. This resulted in a further decrease in sales as their prices went a little above their competitors resulting in their customers moving towards their customers. Even though our organization offered good quality homes, the initial fact of slightly increased pricing led the customers to move towards their competitors right away. Hence, they wanted to reassess their pricing strategy by researching the real estate sales data for twenty years from 2001 to 2020.

**Description of the selected Dataset**

The organization that we have selected is a construction company based out of Connecticut. They are currently running into losses post the covid pandemic because of multiple reasons like higher raw materials cost, delay in the supply chain, pricing strategy not being in line with the market standards, etc., As mentioned by Zaldonis (2023), the Office of Policy and Management maintains a listing of all real estate sales with a sales price of $2,000 or greater that occur between October 1 and September 30 of each year. This is because the annual real estate sales are reported by grand list year - October 1 through September 30 each year. This is the dataset that I will be using in my portfolio project. This dataset is provided by data.ct.gov and is intended for public access and use. For each sale record, the dataset includes fields like town, property address, date of sale, property type (residential, apartment, commercial, industrial or vacant land), sales price, and property assessment. There are around 997213 rows and 14 columns in our dataset.

**Overview of Hypotheses**

From business questions, the researcher attempts to formulate a hypothesis or set of hypotheses to explain what the researcher expects to observe in the reality of the data set. A hypothesis is a proposed explanation of a narrow set of observations, and it represents an informed, theoretical attempt to explain our observations of reality. Hypotheses take two forms: a null hypothesis, given the symbol H0, and an alternative hypothesis, given the symbol HA or H1. The alternative hypothesis states what the research study is attempting to establish. As mentioned by Anderson, et al., (2000), we find that null hypothesis testing is uninformative when no estimates of means, or effect size and their precision are given, and tests of statistical null hypotheses have relatively little utility in science and are not a fundamental aspect of the scientific method. The recommendation is to reduce their use in favor of more informative approaches.

The null and alternative hypothesis for each of our research questions are as follows,

1. What is the average of “Sales Ratio” over a 20-year period?
   1. Null Hypothesis → The average of “Sales Ratio” over a 20-year period is less than 9
   2. Alternative Hypothesis → The average of “Sales Ratio” over a 20-year period is more than 9
2. What is the predicted sales ratio for our construction company for the future three months?
   1. Null Hypothesis → The predicted sales ratio for our construction company is less than 4
   2. Alternative Hypothesis → The predicted sales ratio for our construction company is more than 4

**Literature Review**

As mentioned by Glackin, et. al., (2022), Predicting future demand based on historical sales data, current market fluctuations, and Internet search data, demand management is critical for the strategic and operational planning of a company’s future inventory. This article also mentions that Machine Learning (ML) can support sales managers in making more data-driven, precise, and real-time decisions regarding the optimal use of their resources in sales and marketing. This includes tasks like sentiment analysis, demand management, customer experience management, customer, and market segmentation, etc., This article also covers the disadvantages of ML methods like chances of improvement in the accuracy of forecasting, heavy computing requirements (neural networks), mandatory requirement of quality data representing the actual production data, insecurity to the sales & marketing professionals, etc.,

As mentioned by Hasan, et al., (2021), every step of our research process must be guided by ethical principles. And if the researchers want their studies to be successful, then they must adhere to ethical standards. Authenticity and benevolence are two aspects of truthfulness that are directly related to ethics. It also mentions that researchers must ensure that their methods of data collection and analysis are appropriate and valid and that they do not manipulate or alter the data in any way to convey a different story. They must also be transparent about their methods and ensure their research is replicable. This covers the ethics in all aspects of research like topic selection, research methodology, research results, research discussion, research citation, and that of submitting to a journal.

As part of my Capstone project, I would be using Chi-square test my hypothesis as this is one of the highly used non-parametric test for testing a hypothesis and it tells us how different our observed values were from our predicted values. As mentioned by Franke, et al., (2012), A correct interpretation of the chi-square test or of other statistical procedures is often dependent on factors outside of distributional assumptions and characteristics of the data itself. This article also clarifies that in order to understand the effective use of the chi-square test, or any other statistical test for that matter, we need to have a clear understanding of the assumptions of the test and what is being tested (null hypothesis) in the statistical procedure. More commonly, researchers prefer to interpret the chi-square test of homogeneity by comparing groups across a variable of interest.

From business questions, the researcher attempts to formulate a hypothesis or set of hypotheses to explain what the researcher expects to observe in the reality of the data set. As mentioned by Anderson, et al., (2000), we find that null hypothesis testing is uninformative when no estimates of means, or effect size and their precision are given, and tests of statistical null hypotheses have relatively little utility in science and are not a fundamental aspect of the scientific method. The recommendation is to reduce their use in favor of more informative approaches. This also mentions that Results from null hypothesis testing lead to relatively little increase in understanding and divert attention from the important issues. There are alternatives to the traditional null hypothesis like likelihood ratio, Bayesian methods, etc., that can be used in scenarios where hull hypothesis doesn’t yield required results.

**Capstone Project – Methodology**

The organization that we have selected is a construction company based out of Connecticut. They are currently running into losses post the covid pandemic because of multiple reasons like higher raw materials cost, delay in the supply chain, pricing strategy not being in line with the market standards, etc., As mentioned by Zaldonis (2023), the Office of Policy and Management maintains a listing of all real estate sales with a sales price of $2,000 or greater that occur between October 1 and September 30 of each year. This is because the annual real estate sales are reported by grand list year - October 1 through September 30 each year. This is the dataset that I will be using in my portfolio project. This dataset is provided by data.ct.gov and is intended for public access and use. For each sale record, the dataset includes fields like town, property address, date of sale, property type (residential, apartment, commercial, industrial or vacant land), sales price, and property assessment. There are around 997213 rows and 14 columns in our dataset.

The first variable in our dataset is “Serial number” and it is a discrete variable of type integer. The second variable is “List Year” which is also a discrete variable of type integer. The third variable is "Date Recorded" which is also a discrete variable of type date. The fourth variable is "Town" which is a nominal variable of type string. The fifth variable is “Address” which is also a nominal variable of type string. The sixth variable is "Property Type" which is also a nominal variable of type string. The seventh variable is "Residential Type" which is a binary variable of type string. The eighth variable is "Non Use Code" which is a nominal variable of type string. The ninth variable is "Assessor Remarks" which is also a nominal variable of type string. The tenth variable is "OPM remarks" " which is also a nominal variable of type string. The last variable is “Location” which is also a nominal variable of type string.

As mentioned by Pedamkar (2021), the main pros of quantitative research are that it is quicker and simpler and has a greater sample size. The cons are that the research isn't carried out in the natural surroundings and can be quick and expensive. As mentioned by Smith (2023), qualitative research is based on a constructivist or naturalist approach. In this case, the reality is subjective and seen through the eyes of the participants in the study. Below are the four qualitative research designs that are frequently found in human and social science research,

* Ethnographies, in which a researcher studies a cultural group in a natural setting during a specified period.
* Grounded theory, in which a researcher develops a theory through multiple stages of data collection and compares it with other theories found in the literature.
* Case studies, in which a researcher explores a single phenomenon that occurs during a defined time or activity and collects data.
* Phenomenological studies, in which a researcher examines a human experience through detailed descriptions.

Our selected dataset is from the Office of Policy and Management in Connecticut state, where in they also have the remarks from other parties involved like real estate property Assessor and U.S. Office of Personnel Management (OPM). I assume this data would have been collected by using quantitative methods like feedback surveys and questionnaires. For each real estate sale entry, the remark was provided by OPM and Assessor.

**Tools and Techniques for analyzing the data**

As part of this Capstone project, I had used Chi-square test my hypothesis as this is one of the highly used non-parametric test for testing a hypothesis and it tells us how different our observed values were from our predicted values. As mentioned by Franke, et al., (2012), A correct interpretation of the chi-square test or of other statistical procedures is often dependent on factors outside of distributional assumptions and characteristics of the data itself. We also need to understand that effective use of the chi-square test, or any other statistical test for that matter, is dependent on a clear understanding of the assumptions of the test and what is being tested (null hypothesis) in the statistical procedure. More commonly, researchers prefer to interpret the chi-square test of homogeneity by comparing groups across a variable of interest. If needed, I will also try to use the Two sample t-test as this helps to compare two population averages by comparing two independent samples.

**Security, Privacy and Ethical Considerations**

As mentioned by Glackin, et. al., (2022), Predicting future demand based on historical sales data, current market fluctuations, and Internet search data, demand management is critical for the strategic and operational planning of a company’s future inventory. Our construction company wanted to make more data-driven, precise, and real-time decisions by researching the provided data. The first and foremost benefit that our construction company will achieve from the research is competitive pricing based on the researched data. This will result in an increase in their immediate sales thereby leading to profit. The second benefit is that they will have a solid pricing strategy for the next five years which they can revisit once a year so that they stay close to the realistic home pricing numbers. The dataset that I will be using for my Capstone project is provided by data.ct.gov and is intended for public access and use. For each sale record, the dataset includes fields like town, property address, date of sale, property type (residential, apartment, commercial, industrial, or vacant land), sales price, and property assessment. There are around 997213 rows and 14 columns in our dataset. This data is collected in accordance with Connecticut General Statutes, section 10-261a and 10-261b and we aren’t consuming data which is collected legally by the office of policy and management. Hence it is ethically right to use this data.

As mentioned by Tan, et al., (2013), Accountability of data requires transparency on how data is handled and requires the user to take responsibility for what they do with their personal and company proprietary data by complying with the security standards. Ethical issues related to data privacy and security require policymakers to form regulations and enforce consequences for violations, but policymakers must also be ready to adapt when needed. An effective way to enforce these ethical guidelines within an organization is by asking employees, students, and volunteers to sign a code of conduct. Then, if someone in your profession is not following ethical guidelines, you can appeal to the code of conduct to alter their behavior or to justify their dismissal.

As mentioned by Hasan, et al., (2021), every step of our research process must be guided by ethical principles. And if the researchers want their studies to be successful, then they must adhere to ethical standards. Authenticity and benevolence are two aspects of truthfulness that are directly related to ethics. As mentioned by InnovateMR., (2023), Researchers must ensure that their methods of data collection and analysis are appropriate and valid and that they do not manipulate or alter the data in any way to convey a different story. They must also be transparent about their methods and ensure their research is replicable. This is the main reason behind me choosing to consume my research data from a reputed US government website – catalog.data.gov that only collects data reliable government agencies in an ethical manner. It is a the Data.gov catalog is a consolidated, continually updated catalog of federal datasets. While the OPEN Government Data Act is a requirement for federal agencies, many states, counties, and cities have open datasets and maintain open data websites.

**Capstone Project – Limitation**

The first main limitation is that there are only 11 dimensions in the data and practically we would need more dimensions to enable the organization to analyze their sales better to predict future sales in a more accurate way. Data related tasks like data collection, data cleansing and data analysis are very important for all Analytics projects as they play a key role in deciding the outcome of the data driven analytics that completely depends on the data contained in the data warehouse. So, in our case as well, if this was a real-life project, I would have taken more time in data collection and data cleansing so that I can come up with a valid data set to be used in the analytics. The second limitation is that the remark columns like OPM and Assessor remarks are not filled properly for all houses which makes it difficult to analyze the data properly. This is an example of quantitative research method of collecting feedback surveys. At this point, the main thing to do is to identify this data quality anomaly and fix the same before consuming this data for analytics or machine learning. One option is to use "data imputation" methods to estimate missing/wrong values and use the estimate in place of the true missing/wrong value. As mentioned by Zhang (2016), in data imputation, the missing values are replaced by imputed values. Since imputation is an area of active research, there are numerous methods and packages developed for imputation like regression, indicator methods, etc.,

**Description of Analysis Test results**

1. Import the “Clothing\_Store\_Sales.csv” file in SAS by adding “guessingrows = MAX” so that SAS picks the right length of all columns from the available data.

A screenshot of a computer

Description automatically generated

Figure : Uploading Real\_Estate\_Sales\_2001-2020\_GL.csv file in to SAS

1. Saving the Sales dataset in “WORK” folder

A screenshot of a computer

Description automatically generated

Figure : Saving the Sales dataset in the "WORK" folder

1. Result of the “Descriptive Statistics” SAS command for Real estate sales dataset

A screenshot of a computer

Description automatically generated

Figure : Descriptive Statistics – Results

1. Coding of creating a new table – Sales\_Avg (to save the average of Sales Ratio by year) and the “Series plot” for List\_Year” Vs. “Sales\_Ratio\_Avg”

A screenshot of a computer

Description automatically generated

Figure : Coding for new table - Sales\_Avg and Series Plot

1. Result of the Series Plot

A screenshot of a computer

Description automatically generated

Figure : Result of the Series Plot

1. Create a new aggregated table with List Year, Date Recorded and Sales Ratio Avg and then run the linear regression model for predicting the Sales Ratio values for the four future months

A screenshot of a computer

Description automatically generated

Figure : Result of the Linear Regression

**Result of Hypothesis**

From business questions, the researcher attempts to formulate a hypothesis or set of hypotheses to explain what the researcher expects to observe in the reality of the data set. A hypothesis is a proposed explanation of a narrow set of observations, and it represents an informed, theoretical attempt to explain our observations of reality. Hypotheses take two forms: a null hypothesis, given the symbol H0, and an alternative hypothesis, given the symbol HA or H1. The alternative hypothesis states what the research study is attempting to establish. As mentioned by Anderson, et al., (2000), we find that null hypothesis testing is uninformative when no estimates of means, or effect size and their precision are given, and tests of statistical null hypotheses have relatively little utility in science and are not a fundamental aspect of the scientific method. The recommendation is to reduce their use in favor of more informative approaches.

The null and alternative hypothesis for each of our research questions are as follows,

1. What is the average of “Sales Ratio” over a 20-year period?
   1. Null Hypothesis → The average of “Sales Ratio” over a 20-year period is less than 9
   2. Alternative Hypothesis → The average of “Sales Ratio” over a 20-year period is more than 9

For this research question, alternative hypothesis is TRUE as the average of “Sales Ratio” over a 20-year period is around 10.446 which is higher than 9.

1. What is the predicted sales ratio for our construction company for the future three months?
   1. Null Hypothesis → The predicted sales ratio for our construction company is less than 4.
   2. Alternative Hypothesis → The predicted sales ratio for our construction company is more than 4.

For this research question, we need to reject the alternative hypothesis as the predicted sales ratio for our construction company is around 0.1723 which is less than 4. There is also another important finding that the R2 value in this prediction is 0.0019 which states that our best estimate of the amount of variance explained by the model is that it's miniscule. So, we would need more attributes to come up with a proper prediction in this case.

**Recommendations for further analysis of real estate data**

We had 10 years of real estate sales data with only 11 dimensions and 997213 rows. Because of this we couldn’t properly predict the future three months sales ratio and we need more variables to come up with a proper analysis that is an actual representative of the complete sales data. The second suggestion on my dataset is the remark columns like OPM and Assessor remarks are not filled properly for all houses which makes it difficult to analyze the data properly. This is an example of quantitative research method of collecting feedback surveys. At this point, the main thing to do is to identify this data quality anomaly and fix the same before consuming this data for analytics or machine learning. One option is to use "data imputation" methods to estimate missing/wrong values and use the estimate in place of the true missing/wrong value. As mentioned by Zhang (2016), in data imputation, the missing values are replaced by imputed values. Since imputation is an area of active research, there are numerous methods and packages developed for imputation like regression, indicator methods, etc.,

**Conclusion**

The success of any organization lies in measuring its performance periodically and working towards improving the same (wherever necessary). Our construction company is currently running into losses post the covid pandemic because of multiple reasons like higher raw materials cost, delay in the supply chain, pricing strategy not being in line with the market standards, etc., Our construction company’s first goal is using the provided data to understand the change in the average “Sales Ratio” over a 20-year period. The second goal is to run prediction models on this data for forecasting the future home prices which will help them come up with their pricing roadmap. We succeeded in achieving the first goal of analyzing the change in Sales ratio over the period of 20 years. But regarding the second goal, we had some limitations on the data availability regarding attributes / dimensions and missing data in remarks variables. In the future, once these are fixed, then we will be able to succeed in our second goal as well.

**References**

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: problems, prevalence, and an alternative. The journal of wildlife management, 912-923.

Franke, T. M., Ho, T., & Christie, C. A. (2012). The chi-square test: Often used and more often misinterpreted. American journal of evaluation, 33(3), 448-458.

Glackin, C. E. W., & Adivar, M. (2022). Using the power of machine learning in sales research: process and potential. Journal of Personal Selling & Sales Management, 1–17. <https://doi.org/10.1080/08853134.2022.2128812>

Hasan, N., Rana, R. U., Chowdhury, S., Dola, A. J., & Rony, M. K. K. (2021). Ethical considerations in research. Journal of Nursing Research, Patient Safety and Practise (JNRPSP) 2799-1210, 1(01), 1-4.

John Adams, Hafiz T. A. Khan, & Robert Raeside. (2014). Research Methods for Business and Social Science Students: Vol. Second edition. Sage Publications Pvt. Ltd.

McGaghie, W. C., Bordage, G., & Shea, J. A. (2001). Problem statement, conceptual framework, and research question. Academic medicine, 76(9), 923-924.

Pedamkar, P. (2021, January 19). Types of Research Methodology | Top 10 Types of Research. EDUCBA. https://www.educba.com/types-of-research-methodology/

Smith, T. (2023). Qualitative and quantitative research. Salem Press Encyclopedia.

Tan, Y. S., Ko, R. K., & Holmes, G. (2013, November). Security and data accountability in distributed systems: A provenance survey. In 2013 IEEE 10th International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (pp. 1571-1578). IEEE.

Zaldonis, Pauline. (2023, August 12). Real Estate Sales 2001-2020 GL. <https://catalog.data.gov/dataset/real-estate-sales-2001-2018>

Zhang Z. (2016). Missing data imputation: focusing on single imputation. Annals of translational medicine, 4(1), 9. <https://doi.org/10.3978/j.issn.2305-5839.2015.12.38>