



Competencies

In this project, you will demonstrate your mastery of the following competencies:

- Apply statistical techniques to address research problems
- Perform regression analysis to address an authentic problem

Overview

The purpose of this project is to have you complete all of the steps of a real-world linear regression research project starting with developing a research question, then completing a comprehensive statistical analysis, and ending with summarizing your research conclusions.

Scenario

You have been hired by the D. M. Pan National Real Estate Company to develop a model to predict median housing prices for homes sold in 2019. The CEO of D. M. Pan wants to use this information to help their real estate agents better determine the use of square footage as a benchmark for listing prices on homes. Your task is to provide a report predicting the median housing prices based square footage. To complete this task, use the provided real estate data set for all U.S. home sales as well as national descriptive statistics and graphs provided.

Directions

Using the Project One Template located in the What to Submit section, generate a report including your tables and graphs to determine if the square footage of a house is a good indicator for what the listing price should be. Reference the National Statistics and Graphs document for national comparisons and the Real Estate County Data spreadsheet (both found in the Supporting Materials section) for your statistical analysis.

Note: Present your data in a clearly labeled table and using clearly labeled graphs.

Specifically, include the following in your report:

Introduction

- Describe the report:** Give a brief description of the purpose of your report.
 - Define the question your report is trying to answer.
 - Explain when using linear regression is most appropriate.
 - When using linear regression, what would you expect the scatterplot to look like?
 - Explain the difference between response and predictor variables in a linear regression to justify the selection of variables.

Data Collection

- Sampling the data:** Select a random sample of 50 counties.
 - Identify your response and predictor variables.
- Scatterplot:** Create a scatterplot of your response and predictor variables to ensure they are appropriate for developing a linear model.

Data Analysis

- Histogram:** For your two variables, create histograms.
- Summary statistics:** For your two variables, create a table to show the mean, median, and standard deviation.
- Interpret the graphs and statistics:**

- a. Based on your graphs and sample statistics, interpret the center, spread, shape, and any unusual characteristic (outliers, gaps, etc.) for the two variables.
- b. Compare and contrast the shape, center, spread, and any unusual characteristic for your sample of house sales with the national population. Is your sample representative of national housing market sales?

Develop Your Regression Model

- A. **Scatterplot:** Provide a graph of the scatterplot of the data with a line of best fit.
 - a. Explain if a regression model is appropriate to develop based on your scatterplot.
- B. **Discuss associations:** Based on the scatterplot, discuss the association (direction, strength, form) in the context of your model.
 - b. Identify any possible outliers or influential points and discuss their effect on the correlation.
 - c. Discuss keeping or removing outlier data points and what impact your decision would have on your model.
- C. **Find r :** Find the correlation coefficient (r).
 - a. Explain how the r value you calculated supports what you noticed in your scatterplot.

Determine the Line of Best Fit. Clearly define your variables. Find and interpret the regression equation. Assess the strength of the model.

- A. **Regression equation:** Write the regression equation (i.e., line of best fit) and clearly define your variables.
- B. **Interpret regression equation:** Interpret the slope and intercept in context.
- C. **Strength of the equation:** Provide and interpret R -squared.
 - a. Determine the strength of the linear regression equation you developed.
- D. **Use regression equation to make predictions:** Use your regression equation to predict how much you should list your home for based on the square footage of your home.

Conclusions

- A. **Summarize findings:** In one paragraph, summarize your findings in clear and concise plain language for the CEO to understand. Summarize your results.
 - a. Did you see the results you expected, or was anything different from your expectations or experiences?
 - i. What changes could support different results, or help to solve a different problem?
 - ii. Provide at least one question that would be interesting for follow-up research.

What to Submit

To complete this project, you must submit the following:

Project One Template: Use this template to structure your report, and submit the finished version as a Word document.

Supporting Materials

The following resources may help support your work on the project:

Document: [National Statistics and Graphs](#)

Use this data for input in your project report.

Spreadsheet: [Real Estate County Data](#)

Use this data for input in your project report.

Tutorial: [Downloading Office 365 Programs](#)

Use this tutorial for support with Office 365 programs.

Use these tutorials for support with the Excel functions you will use in the project:

- **Tutorial:** [Random Sampling in Excel](#)

- Tutorial: [Scatterplots in Excel](#)
- Tutorial: [Descriptive Statistics in Excel](#)
- Tutorial: [Creating Histograms in Excel](#)

Project One Rubric

Criteria	Exemplary	Proficient	Needs Improvement	Not Evident	Value
Introduction: Describe the Report	Exceeds proficiency in an exceptionally clear manner (100%)	Defines the question the report is trying to answer, including when using linear regression is most appropriate, what the scatterplot will look like, and the difference between response and predictor variables in a linear regression to justify the selection of variables (85%)	Shows progress toward proficiency, but with errors or omissions; areas for improvement may include inaccurately defining the question, the appropriateness and justification of the linear regression model or the selection of variables, or introduction lacking essential detail and clarity (55%)	Does not attempt criterion (0%)	10
Data Collection: Sampling the Data	N/A	Selects a random sample of 50 counties and identifies the response and predictor variables (100%)	Shows progress toward proficiency, but with errors or omissions; areas for improvement may include inaccurate selection of random sample or inaccurate or unclear selection of response and predictor values (55%)	Does not attempt criterion (0%)	5
Data Collection: Scatterplot	N/A	Creates a scatterplot of the response and predictor variables to ensure they are appropriate for developing a linear model (100%)	Shows progress toward proficiency, but with errors or omissions; areas for improvement may include inaccurate scatterplot representation of the information or inaccurate or unclear determination of response and predictor variables (55%)	Does not attempt criterion (0%)	5

Criteria	Exemplary	Proficient	Needs Improvement	Not Evident	Value
Data Analysis: Histogram	N/A	Creates histograms for the two variables (100%)	Shows progress toward proficiency, but with errors or omissions; areas for improvement may include histograms that are created incorrectly or are inaccurate (55%)	Does not attempt criterion (0%)	5
Data Analysis: Summary Statistics	N/A	Creates a table to show the mean, median, and standard deviation for two variables (100%)	Shows progress toward proficiency, but with errors or omissions; areas for improvement may include table showing mean, median and standard deviation that are inaccurate or created incorrectly (55%)	Does not attempt criterion (0%)	5
Data Analysis: Interpret Graphs and Statistics	N/A	Interprets the graphs and statistics center, spread, shape, and any unusual characteristic (outliers, gaps, etc.) for the two variables based on the graphs and sample statistics and compares with national housing market sales (100%)	Shows progress toward proficiency, but with errors or omissions; areas for improvement may include inaccurate or cursory interpretation of the characteristics of the graph and statistics or inaccurate or cursory comparison with the national market (55%)	Does not attempt criterion (0%)	5
Develop Regression Model: Scatterplot	N/A	Provides a graph of the scatterplot of the data with a line of best fit; explains if a regression model is appropriate to develop based on the scatterplot (100%)	Shows progress toward proficiency, but with errors or omissions; areas for improvement may include Inaccurate scatterplot or line of best fit or explanation of regression model appropriateness that is inaccurate or cursory (55%)	Does not attempt criterion (0%)	5

Criteria	Exemplary	Proficient	Needs Improvement	Not Evident	Value
Develop Regression Model: Discuss Associations	Exceeds proficiency in an exceptionally clear manner (100%)	Discusses the association in the context of the model based on scatterplot, includes possible outliers or influential points, discusses effect on correlation, and discusses impact of keeping or removing outliers (85%)	Shows progress toward proficiency, but with errors or omissions; areas for improvement may include discussion of association in the context of the scatterplot, possible outliers, influential points and impact on correlation, or impacts of keeping or removing outliers that is inaccurate or cursory (55%)	Does not attempt criterion (0%)	10
Develop Regression Model: Find r	Exceeds proficiency in an exceptionally clear manner (100%)	Finds the correlation coefficient (r) and explains how the calculated r value supports what was noticed in the scatterplot (85%)	Shows progress toward proficiency, but with errors or omissions; areas for improvement may include inaccurate calculation for r or explanation of how the r value supports the scatterplot that is inaccurate or cursory (55%)	Does not attempt criterion (0%)	10
Determine Line of Best Fit: Regression Equation	N/A	Writes the regression equation and clearly defines variables (100%)	Shows progress toward proficiency, but with errors or omissions; areas for improvement may include regression equation that is written inaccurately or variables that are not clearly defined (55%)	Does not attempt criterion (0%)	5
Determine Line of Best Fit: Interpret Regression Equation	Exceeds proficiency in an exceptionally clear manner (100%)	Interprets the slope and intercept in context (85%)	Shows progress toward proficiency, but with errors or omissions; areas for improvement may include inaccurate interpretation of the slope and intercept (55%)	Does not attempt criterion (0%)	10

Criteria	Exemplary	Proficient	Needs Improvement	Not Evident	Value
Determine Line of Best Fit: Strength of the Equation	N/A	Provides and interprets <i>R-squared</i> , determining the strength of the linear regression equation (100%)	Shows progress toward proficiency, but with errors or omissions; areas for improvement may include inaccuracies in interpretation of <i>R-squared</i> or the determined strength of the regression equation (55%)	Does not attempt criterion (0%)	5
Determine Line of Best Fit: Use Regression Equation to Make Predictions	N/A	Uses a regression equation to predict how much you should list your home for based on the square footage of your home (100%)	Shows progress toward proficiency, but with errors or omissions; areas for improvement may include misuse of regression equation or inaccurate prediction based on provided information (55%)	Does not attempt criterion (0%)	5
Conclusion: Summarize Findings	Exceeds proficiency in an exceptionally clear manner (100%)	Summarizes findings and results in clear and concise plain language, includes whether the results were expected, changes that could support different results or that would help to solve a different problem; Includes a question for follow-up research (85%)	Shows progress toward proficiency, but with errors or omissions; areas for improvement may include inaccurately summarizing findings or results or summary that is cursory or missing required elements (55%)	Does not attempt criterion (0%)	10
Articulation of Response	Exceeds proficiency in an exceptionally clear, insightful, sophisticated, or creative manner (100%)	Clearly conveys meaning with correct grammar, sentence structure, and spelling, demonstrating an understanding of audience and purpose (85%)	Shows progress toward proficiency, but with errors in grammar, sentence structure, and spelling, negatively impacting readability (55%)	Submission has critical errors in grammar, sentence structure, and spelling, preventing understanding of ideas (0%)	5
Total:					100%