**Assignment 3**

Use appropriate **Excel** Data Analysis Tools to perform the analyses specified below and answer the given questions using the two data sets provided. Save any analysis results and other work in the Excel files, but also copy and paste charts or tables from Excel into this Word document as requested. For some questions, there is a formula template provided that you can modify using the Equation Editor in Word (The templates are guides – your final equations and notation may vary from what is given.). Upload this completed worksheet, along with your two modified Excel files, by the time and date specified in Canvas.

Scoring for assignments will be based not only on the accuracy of your data analysis, but also on the organization, clarity, and quality of your written answers and this completed summary.

**Case 1: Analyzing factors that may influence the cost of apartment rentals.** The file Dorchester.xlsx contains a sample set of data from apartment rentals in the Dorchester part of Boston. Provided for each rental are 1) the monthly rent ($US) for the apartment, 2) the number of bedrooms, 3) the number of bathrooms, and 4) the total size of the apartment (in square feet). We are interested in analyzing the relationship between Rent and the other three variables.

**A.** Create a correlation matrix for the entire data set and paste a copy of it below. Based solely on this analysis, which **single** variable would create the best simple linear regression equation with *Rent* as the dependent variable? Justify your choice.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Rent* | *Beds* | *Baths* | *Size* |
| Rent | 1 |  |  |  |
| Beds | 0.86880926 | 1 |  |  |
| Baths | 0.72677604 | 0.63835931 | 1 |  |
| Size | 0.83930727 | 0.81553167 | 0.72277342 | 1 |

Baths would be the single best variable for creating the best sample linear regression equation with Rent as the dependent variable as it has the highest value when the dependent variable is Rent (0.86880926).

**B.** Create a **simple linear regression equation** using Rent (y) and the best independent variable (x) that you identified in Step A. Paste a copy of the Regression Statistics and Coefficients tables below, and then provide the final regression equation (the template below can be edited to reflect the correct variables and coefficients).

|  |  |
| --- | --- |
| *Regression Statistics* | |
| Multiple R | 0.86880926 |
| R Square | 0.75482954 |
| Adjusted R Square | 0.74837768 |
| Standard Error | 212.179989 |
| Observations | 40 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 1359.47391 | 86.5911537 | 15.6999168 | 3.3679E-18 | 1184.17928 | 1534.76854 | 1184.17928 | 1534.76854 |
| Beds | 388.067906 | 35.8777799 | 10.8163857 | 3.6872E-13 | 315.437137 | 460.698674 | 315.437137 | 460.698674 |

**C.** State the value and interpret the meaning of the coefficient of multiple determination (*r2*) of your regression equation in Step B. A template is provided for *r2*.

Since this value is close to 1 (greater than 0.70), that means there is a strong linear relationship between the number of beds and the rent.

**D.** Now create a multiple regression model for Rent using **all three** of the other variables. Paste the Regression Statistics and Coefficients tables below and also provide the final regression equation.

|  |  |
| --- | --- |
| *Regression Statistics* | |
| Multiple R | 0.90794324 |
| R Square | 0.82436092 |
| Adjusted R Square | 0.80972433 |
| Standard Error | 184.510663 |
| Observations | 40 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 1296.92549 | 79.6766792 | 16.2773537 | 3.5268E-18 | 1135.33369 | 1458.51728 | 1135.33369 | 1458.51728 |
| Beds | 232.909982 | 54.3190968 | 4.28781029 | 0.0001291 | 122.745748 | 343.074217 | 122.745748 | 343.074217 |
| Baths | 102.657696 | 52.7565071 | 1.94587743 | 0.05950873 | -4.3374594 | 209.652852 | -4.3374594 | 209.652852 |
| Size | 0.17738678 | 0.08871633 | 1.99948284 | 0.05314686 | -0.0025383 | 0.35731184 | -0.0025383 | 0.35731184 |

**E.** State the value and interpret the meaning of adjusted *r2* for this multivariate model. A template is provided.

Since this value is close to 1, that means there is a strong linear relationship between the number of beds, number of baths, and size to the rent.

**F.** If you needed to choose one of these two models to predict rent for an apartment, would it be the one from Step B or Step D? Why? Justify your decision using the appropriate statistics for each model.

If needed, it would be best to choose the model from Step D since the adjusted value of the model in Step D is greater (and thus closer to 1) than the model in Step B meaning that there is a stronger linear relationship in Step D.

**Case 2: Estimating future revenues from past results.** The file Revenue\_B.xlsx contains 8 years of revenue data for a business. The data is in U.S. dollars and is broken down by quarter for each year. We want to estimate revenue for 2009 based on this historical data. For this problem, you may need to create additional variables to perform your analysis.

**A.** Create a line chart of Revenue over time. Be sure to add an appropriate title and axis labels. Paste a copy of the chart below. Describe the overall trend, as well as any potential seasonality components, based on a visual analysis of the data in this chart. Explain how the data supports your analysis of trend and seasonality.

The overall trend shows a decrease in the revenue over time between the years of 2001 and 2008 and there seems to be a trend of a general decrease in revenue in the second quarter, which is backed up by the data as every year there is a drop in the revenue value between quarter 1 and quarter 2

**B.** Estimate a linear trend model for Revenue. Paste a copy of just the Regression Statistics and Coefficients tables below. Then provide the final forecast equation (you can use the template below). Be sure to clearly define your variables. How well does the model seem to work? Provide support for this answer.

|  |  |
| --- | --- |
| *Regression Statistics* | |
| Multiple R | 0.89420678 |
| R Square | 0.79960577 |
| Adjusted R Square | 0.79292596 |
| Standard Error | 65776.0676 |
| Observations | 32 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 1441614.47 | 23811.335 | 60.5432021 | 6.3559E-33 | 1392985.24 | 1490243.7 | 1392985.24 | 1490243.7 |
| t | -13778.492 | 1259.3479 | -10.940974 | 5.4108E-12 | -16350.424 | -11206.561 | -16350.424 | -11206.561 |

Where t = (year – 2001) \* 4 + Quarter (or just incrementing from 1 to 32 for all the values in the table)

This model seems to be strong since it has an Adjusted R Squared value of 0.79292596 which is greater than 0.7 and close to 1, meaning there is a strong linear relationship between t and the revenue.

**C.** Now estimate a linear trend model for Revenue which **also** accounts for seasonality (through the use of dummy variables). Paste a copy of the Regression Statistics and Coefficients tables below, and then provide the final forecast equation (rename the variables used in the template as appropriate).

|  |  |
| --- | --- |
| *Regression Statistics* | |
| Multiple R | 0.93517513 |
| R Square | 0.87455253 |
| Adjusted R Square | 0.85596772 |
| Standard Error | 54857.3727 |
| Observations | 32 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 1470173 | 27182.6842 | 54.0849088 | 4.577E-29 | 1414398.73 | 1525947.26 | 1414398.73 | 1525947.26 |
| t | -13716.361 | 1058.08455 | -12.963388 | 4.156E-13 | -15887.371 | -11545.351 | -15887.371 | -11545.351 |
| q1 | 17434.6674 | 27611.7497 | 0.63142204 | 0.53307437 | -39219.963 | 74089.2981 | -39219.963 | 74089.2981 |
| q2 | -72849.847 | 27510.1982 | -2.6481033 | 0.01335183 | -129296.11 | -16403.582 | -129296.11 | -16403.582 |
| q3 | -62919.611 | 27449.087 | -2.2922296 | 0.02991358 | -119240.49 | -6598.7365 | -119240.49 | -6598.7365 |

Where q1, q2, and q3 represent quarters 1, 2, and 3 respectively and t represents what it does in Step B.

**D.** Does the model from Step B or Step C work better to forecast Revenue? Explain why, based on the statistics generated by each model.

The model from step C works best since the adjusted value of the model in Step C is greater (and thus closer to 1) than the model in Step B meaning that there is a stronger linear relationship in Step C.

**E.** Use the better model (as chosen in Step D) to forecast revenue for 2009.

|  |  |  |
| --- | --- | --- |
| Year | Quarter | Revenue ($US) |
| 2009 | 1 | 1034967.75 |
| 2009 | 2 | 930966.879 |
| 2009 | 3 | 927180.754 |
| 2009 | 4 | 976384.004 |