WISCONSIN SCHOOL OF BUSINESS
UNIVERSITY OF WISCONSIN–MADISON

TOGETHER FORWARD®

Business Analytics I – Case II

General Instructions:

1. You must complete this project individually. Only work with the sample assigned to you (not the full dataset) for all parts of this case.
2. Read the case ("PhillyCycle Case.pdf") before tackling these instructions.
3. After reading the case, proceed to the questions given in this assignment.
4. There are three submissions: Part I, Part II, and Part III. For each submission, you need to submit your responses along with your Excel file on Canvas. Make sure this file is easy-to-follow and includes the necessary formulas, charts, pivot tables, etc., to support your work. For Part III, please follow the provided template "PhillyCycle Case Template.docx".
5. Use the following z-score values: z-score for a 99% CI = 2.58, z-score for a 95% CI = 1.96, z-score for a 90% CI = 1.64.
6. Round all your submitted responses to 2 decimal points, not 2 significant digits. E.g., 14.075 is entered as 14.08. Round only at the last step of your calculations. Incorrect rounding will be penalized, so pay attention to this detail. Report percentages and proportions as their decimal equivalent (i.e., do not use a percent sign, so 23.45% would be entered as 0.23).

Case Objective:

Help PhillyCycle analyze its customer bike rental behavior, focusing on the factors associated with bike rentals that could be relevant to expansion strategies.

Grading:

Your case will be graded on 100 points using the following criteria and weights:

Data Preparation

Checkpoint Certification due by 11:59pm CST on Friday, 11/12	10 points

Your Analyses

Part I due by 11:59pm CST on Friday, 11/19	30 points
Part II due by 11:59pm CST on Friday, 12/10	40 points

Presentation of your analyses

Part III (Case Report) due by 11:59pm CST on Tuesday, 12/14	20 points

Data preparation

PhillyCycle provided two years of their bike rental data, showing bike rentals initiated by hour, for 2011 and 2012. Each record (row of data) represents **one hour** of bike rentals. You will analyze a large sample of these data. Download the data "PhillyCycle Data.xlsx" from Canvas and check that you have the following variables:

- **record_number** - unique record number for each observation
- **datetime** - hourly date + timestamp
- **season** - 1 = Winter, 2 = Spring, 3 = Summer, 4 = Fall
- **holiday** - whether the day is considered a holiday
- **weather** - 1 = Clear to Partly Cloudy, 2 = Mist, 3 = Precipitation (rain or snow)
- **temp** - temperature in Celsius
- **humidity** - relative humidity (0% to 100%)
- **windspeed** - wind speed in miles per hour
- **casual** - number of non-registered user rentals initiated in a given hour
- **registered** - number of registered user rentals initiated in a given hour
- **all_users** - number of total rentals initiated in a given hour

The Analytics Team at PhillyCycle indicated that their records should be fairly clean. In keeping with good practice, however, you'll want to look over the data for any obvious data problems. Specifically:

i. Make a copy of the worksheet and rename the sheet containing the original dataset "OriginalData" and the dataset you are about to process "ProcessedData".

ii. Create your sample. In the “ProcessedData” sheet, delete the rows of data outside the range of the record numbers assigned to your teams.

iii. Remove any duplicate entries from the data (the record number is always unique, so remember to unselect this column when checking for duplicates).

iv. Check all other variables using filters to ensure that your data make sense; this is an important preliminary step in any data analysis. Specifically, PhillyCycle has informed us that there are some data entry errors in the temperature column, so you should remove any row of data for which the temperature is greater than 50 Celsius (this equals 122 Fahrenheit.)

The following steps will help you prepare some columns of data to use in regression analysis. These steps derive from knowing the various ways in which you plan to look at your data. For example, if you suspect that season or day of the week might be related to rentals, you will want columns of data showing the season or day of the week, which you can derive from the date information. These instructions for data preparation are not meant to exhaust the possibilities for data exploration, but they provide you a good start.

v. Create additional variables about the timing of rentals. Add six new columns in your “ProcessedData” worksheet labeled “Year”, “Month”, “DayofWeek”, “Hour”, “Weekend”, and “WeekNumber”. These variables will denote, respectively, the year (2011 or 2012), the month of the year (1-12), the day of the week (1-7), the hour of the day (0-23), if the rentals occurred on a weekend (0-1), the week number (1-53). The functions YEAR, MONTH, WEEKDAY, HOUR, WEEKNUM, and IF combined with OR, could be used to create these variables.

A good idea is to check some of the dates on a calendar to make sure you are coding your weekdays properly. To make sure that 1 corresponds to Monday, use the following Excel formula: “=WEEKDAY(B2,**2**)” (if the variable “datetime” is in column B). Also, for the week number use “=WEEKNUM(B2,**1**)”.

vi. Using “IF” statements create two categorical variables for the weather. Add two columns to your “ProcessedData” worksheet and name them “Mist” and “Precipitation”. Populate these columns with 1 if the weather pattern falls into the category and 0 otherwise. (These variables are being created for the regression analysis, and the reason there are not three new columns is that in regression, when Mist and Precipitation are both zero, the regression model will be describing the default case, Clear to Partly Cloudy.)

vii. Use the “season” variable (in column C) and “IF” statements to create three categorical variables (i.e., three new columns) for the seasons: Spring, Summer, and Fall. Populate these columns with 1 if the observation occurred during the season and 0 otherwise. In addition, use “IF” statements to create another categorical variable named “Year 2012”. Populate this column with 1 if the observation is in year 2012 and 0 otherwise. These variables will be used for

regression analysis, and you do not need to create a categorical variable for Winter or year 2011 as they will be the default category for seasons and years, respectively.

viii. Create one new column for converting temperature from Celsius to Fahrenheit Excel has a function, CONVERT, which you can use, or you may implement the following formula: Fahrenheit $= \left(\text{Celsius} \times \frac{9}{5}\right) + 32$. Name this variable "Temp (F)", and for the remainder of the case, use Fahrenheit as your measure of temperature.

Formatting your Excel file (suggested): Create and label a new worksheet for each question. This sheet should contain all the necessary work, whether it be formulas, pivot tables, charts, etc., for that question and its sub-parts. You can build on the same file throughout the case and "hide" worksheets if you get overwhelmed with scrolling.

Checkpoint [10 points]

After you have completed all the previous steps, please answer the following questions:

a) the starting and ending record number of your dataset;
b) the number of observations in your dataset (remembering to omit header rows);
c) the average number of total rentals initiated in an hour for year 2012;
d) the number of hourly observations that occurred on a Monday;
e) the max of temperature in degrees Fahrenheit;
f) the standard deviation of number of casual rentals initiated in an hour on Tuesdays;
g) the lower and upper bound of a 95% confidence interval for the average number of registered rentals initiated in an hour for the two years (remember to use the z-scores that are listed on the first page of these instructions)

Your answers must exactly match (to 4 decimal places) the ones posted on Canvas. The PDF on Canvas is arranged by sample number. Your sample number is in Canvas with your assigned record numbers.

There will be a Canvas survey asking if your Checkpoint answers exactly match the ones on Canvas. Be sure you can answer "Yes" to this survey.

As with the first case, you must match before moving on to the rest of the case. If you do not match, then your cleaned data set is incorrect, and you will lose many points throughout the rest of the case.

Part I: Summary Analysis

1. [3 points] Begin by summarizing the number of rentals by casual users. Use Excel to generate the following summary statistics and report your responses:
    a) Average
    b) Median
    c) Minimum
    d) Maximum
    e) First quartile (use QUARTILE.EXC)
    f) Third quartile (use QUARTILE.EXC)
    g) Standard deviation (use STDEV.S)

2. [3 points] Begin by summarizing the number of rentals by registered users. Use Excel to generate the following summary statistics and report your responses:
    a) Average
    b) Median
    c) Minimum
    d) Maximum
    e) First quartile (use QUARTILE.EXC)
    f) Third quartile (use QUARTILE.EXC)
    g) Standard deviation (use STDEV.S)

3. [4 points] We also want to examine the relative importance of casual vs. registered user segment by looking at the proportion of casual rentals out of total rentals. Report:
    a) the total number of casual rentals that are in your sample;
    b) the total number of registered rentals in your sample;
    c) Based on your calculations in (a) and (b), what proportion of all rentals are casual rentals?
    d) Using only year 2011 data, what proportion of all rentals are casual rentals?

4. [3 points] Now turn your attention to the time and seasonal effects. We first look at the average number of hourly rentals by each user group (all users, casual and registered) in each season by using a pivot table to summarize it. Report
    a) the average number of hourly rentals by all users in spring;
    b) the average number of hourly rentals by casual users in summer;
    c) the average number of hourly rentals by registered users in fall;

5. [4 points] Use pivot tables to obtain the total number of rentals by each user group (all users, causal and registered) in each month of 2011 and each month of 2012 (you should have 24 numbers for each group). Also summarize the total number of rentals by all users in each month (combining 2011 and 2012 this time, you should have 12 numbers). Report:
    a) the total number of rentals by casual users in Aug 2011;
    b) the total number of rentals by registered users in Feb 2012;

c) the total number of rentals by all users in Aug 2012;
d) the total number of rentals by all users in Feb (2011 and 2012 combined).

6. [3 points] Use a pivot table to summarize the average hourly rentals by each user group (all users, casual users, and registered users) over the days of the week (you should have 7 numbers for each group). Report:
    a) the average number of hourly rentals by all users on Mondays;
    b) the average number of hourly rentals by casual users on Fridays;
    c) the average number of hourly rentals by registered users on Saturdays.

7. [6 points] Use pivot tables to summarize the average of rentals by all users, the average of rentals by casual users, and the average of rentals by registered users by hour in the day, over weekday vs. weekend. For EACH user group (all users, casual and registered), your pivot table should show the hourly average rental volume by weekday and weekend (48 numbers for each group: 24 for weekday, 24 for weekend). Report:
    a) the average number of rentals by all users at 9 am on weekdays;
    b) the average number of rentals by all users at 9 am on weekends;
    c) the average number of rentals by casual users at 11 am on weekdays;
    d) the average number of rentals by casual users at 11 am on weekends;
    e) the average number of rentals by registered users at 1 pm on weekdays;
    f) the average number of rentals by registered users at 1 pm on weekends;

8. [4 points] Use "If statements" or the histogram option from Excel's "Analysis Toolpak" to answer the following questions. Use these bins: 0-50,51-100,101-150, …, 351-400, and >=401. Report:
    a) The number of hourly rental values by the casual users that is in the second bin (51-100);
    b) The number of hourly rental values for the registered users that is in the first bin (0-50).

(Before you submit, please make sure all your numbers are rounded to 2 decimal points.)

Part II: Regression Analysis

1. [3 points] We start by providing some confidence intervals of the average number of rentals for each user group. Construct and report:
    a) a 95% confidence interval of the average number of hourly rentals by all users;
    b) a 95% confidence interval of the average number of hourly rentals by casual users;
    c) a 99% confidence interval of the average number of hourly rentals by registered users;

2. [8 points] Conduct hypothesis tests, all at 5% level of significance, to determine if there is evidence of different hourly rental volume on weekends compared to

weekdays for all users, for rentals by registered users, and for rentals by casual users.
Use the Confidence Interval formula for the different in means:

$$(\overline{x_1} - \overline{x_2}) \pm z - score * \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Where $\overline{x_1}$ is weekend and $\overline{x_2}$ is weekday (

Though if you were doing this for your own work, you could choose either variable, being sure to interpret the results correctly for whichever variable you choose for $\overline{x_1}$ and $\overline{x_2}$.

To help you prepare the statistics needed for the tests, use pivot tables to summarize the following information and report:

a) the sample mean, sample standard deviations and sample sizes (i.e., how many hours are sampled in your data set) for hourly rentals by all users on weekend and weekdays;
b) the sample mean, sample standard deviations and sample sizes for hourly rentals by casual users on weekend and weekdays;
c) the sample mean, sample standard deviations and sample sizes for hourly rentals by registered users on weekend and weekdays;
d) your conclusion of testing on whether the average number of rentals by all users on weekends is significantly different from the rentals by all users on weekdays;
e) your conclusion of testing on whether the average number of rentals by casual users on weekends is significantly different from the rentals by casual users on weekdays;
f) your conclusion (reject or fail to reject) of testing on whether the average number of rentals by registered users on weekends is significantly different from the rentals by registered users on weekdays. Remember, if the interval contains 0, then the two variables could have the same average.

Analyze rentals by all users, rentals by registered users, and rentals by casual users, by conducting regressions as described below:

3. [3 points] Perform a regression analysis to explain how rentals by all users (Y) change with temperature (remember, always in Fahrenheit).
   a) Report the coefficient of temperature;
   b) What proportion of the variation in rentals is explained by variation in temperature?
   c) Report the correlation between the two variables (rentals by all users and temperature).

4. [6 points] Conduct three more regressions to explain the relationships between the following variables:
   a) Regress all users (Y) on temperature and weekend. Report the coefficient of weekend in this model;
   b) Is the weekend variable statistically significant (at 10% level) in the regression model in (a)?
   c) Regress casual users (Y) on temperature and weekend. Report the coefficient of weekend in this model.
   d) Is the weekend variable statistically significant (at 5% level) in the regression model in (c)?
   e) Regress registered users (Y) on temperature and weekend. Report the coefficient of weekend in this model.
   f) Is the weekend variable statistically significant (at 1% level) in the regression model in (e)?

5. [12 points] Run a multiple regression predicting "all_users" (Y) with variables for: temperature (in Fahrenheit), windspeed, humidity, and dummy (categorical) variables for season, year, weekend, and weather. Be careful to choose the columns to include in your regression in Excel. Now, analyze your regression results to answer the following questions:
   a) Report the coefficient of Mist in your model;
   b) Is the Fall season a statistically significant predictor of rentals at 1% level?
   c) Is the temperature a statistically significant predictor of rentals at 5% level?
   d) All else equal, how many more/less rentals are in Summer compared to Winter? (If there are less rentals, enter a negative number on Canvas)
   e) What is the effect of a one-degree (Fahrenheit) increase in temperature on rentals? (Again, enter a positive number for more rentals and a negative number for less rentals)
   f) All else equal, how many more/less rentals are expected on a clear day versus a day with precipitation (rain or snow)? (enter in the same fashion as part (e))
   g) Report the expected number of rentals for a partly cloudy, Summer, weekday hour in 2012 with the following features: temperature=70 (Fahrenheit), windspeed=5, and humidity=60;
   h) According to the regression model, how much of the variation (in %) in bike rentals is explained by the variation in <u>all</u> the independent variables?
   i) What is the effect of a one-degree Celsius increase in temperature on rentals? <u>This question may be answered without running a new regression</u>. (Again, enter a positive number for more rentals and a negative number for less rentals)

6. [4 points] Use a pivot table to calculate the average rentals and temperature (in Fahrenheit) for all users in each week of the year. Use the WeekNumber variable; your pivot table should have 53 rows, one for each week. Report:
   a) The correlation between the average weekly rentals and average weekly temperature.
   b) The R-squared value from the trendline if you create a scatterplot of the weekly average rentals and temperature.

7. [4 points] Answer the following two questions using regression analysis:
   a) What percent of variation in bike rentals for all users is explained by the variation in temperature, season, and year?
   b) It may become too costly for PhillyCycle to collect both windspeed and humidity information. Suppose the company is chiefly interested in predicting the variable "all_users" using a regression model that includes: temperature (in Fahrenheit) and dummy (categorical) variables for season, year, weekend, and weather. Which of the two variables (windspeed or humidity) is more useful addition to this regression model?

      We can create two regression models, one with windspeed (and not humidity) and the rest of the variables, and the other with humidity (and not windspeed) and the rest of the variables. Remember, the higher the R^2, the stronger the linear model fits the data.

(Before you submit, please make sure all your numbers are rounded to 2 decimal points.)

Part III: Case Report

A primary goal of Business Analytics I is to illustrate how analytical techniques can be used to address business problems. Can you correctly and adequately interpret the results of your analyses considering the business situation? Can you provide the executives at PhillyCycle with meaningful interpretations of the analyses you have conducted – ones that will help them make decisions about how to go forward? The case report is the key deliverable of this process. Your report must follow the guidelines provided in lecture and readings for creating good charts and communicating analyses effectively. It should be succinct, polished, and free of typos. There is no page limit requirement, but follow the structure provided in the template.

1. Produce an executive summary of your analyses (250 to 500 words.) This summary should cover all the highlights of your analysis and provide your recommendations.

2. Graph the data and study the patterns for insights that would be relevant to any expansion plans that PhillyCycle might pursue. In particular, look for differences between the two customer groups (registered and casual).

   a) **(Similar to Part I Q1 and Q2)** Choose three of the summary measures (for each of the two segments) to communicate with PhillyCycle and explain your choices. For example, if you choose to include the average and not the median, explain why. Broadly, what do you learn about bike rentals from your sample based on these measures? Explain in 2-3 sentences.

b) **(Similar to Part I Q3)** Use a pivot table to produce a visualization (your choice) showing the proportions of all bike rentals comprised of casual and registered user segments (2011 and 2012 combined). Interpret the chart in 1-2 sentences.

c) **(Similar to Part I Q4)** Produce a column chart summarizing the average of rentals by all users by season (2011 and 2012 combined). Using 2-3 sentences, note the differences and similarities of demand patterns by season.

d) **(Similar to Part I Q5)** Create a column chart of the monthly totals for all users with 2011 and 2012 columns next to each other, e.g., Jan 2011 next to Jan 2012 and so on. Use 2-3 sentences to interpret the chart.

e) **(Similar to Part I Q6)** Create a line chart to depict the average rental volume by all users, casual users, and registered users across the days of week (your chart should have three lines total). Comment on the differences between the two user segments (casual vs. registered). Is there a pattern in the number of rentals across the days of the week? Are the patterns different for casual, registered, and all users?

f) **(Similar to Part I Q7)** Create two line charts, one for weekday and one for weekend, of the hourly average number of rentals in each hour by each user segment. Make sure there are two lines in each chart, one for the registered renters and one for the casual renters. Is there a pattern for the hourly number of rentals? Are the patterns different for casual, registered rentals?

g) **(Similar to Part I Q8)** Create two histograms showing the distribution of hourly weekday rentals by casual and by registered users. The y-axis unit must be in percentages (not frequencies). Comment on the plots.

h) **(Similar to Part II Q6)** Produce the scatterplot referenced in Q6b. Add the trendline and display the R-squared and regression equation. Interpret the R-squared value and the regression equation. Do the two variables appear to be strongly correlated?

3. **(Similar to Part II Q2)** Describe the test conclusions about the difference between weekday and weekend rentals for each user group (all users, casual and registered). Interpret your results.

4. **(Similar to Part II Q5)**
   a) Describe the regression analysis you have done in Part II Q5.
   b) Show the regression equation. Use the following format (example for formatting only):

      *All Rentals = a + b1*Temperature + b2*Summer + b3*Year2012.*

c) Show your table of regression results.
d) Interpret your regression results. Remember to identify the significant factors and interpret R-squared.

5. Based on your analysis, summarize your major insights, and provide recommendations that PhillyCycle could follow to expand their rental business. You might not have final solutions, but there is great leverage in isolating factors associated with a challenge so that further steps can be taken. What could your analysis mean in the context of the business challenge? What are the benefits and drawbacks of your analyses?

6. Provide a high-level summary (up to 200 words) of the data preparation process you undertook. Are there any concerns you have with the data and the way you have analyzed it? Provide the executives a sense of data quality, too: were many duplicates or erroneous records in your data set?

7. **(Elevator Pitch as an Appendix)** Choose three of the plots/tables that you already created in the case as your "elevator charts," and number them in the order of importance. That is, prioritize your charts so that the first one is what you find most compelling, and would share in an "elevator pitch" to the executives, and briefly explain both your choice of charts and your rationale for their prioritization. Reproduce the charts that you chose in this section.