**MEM 490, Winter 2018 - Managerial Analytics**

**Due: 6:30pm, Monday, March 19th**

**Electronic Submission Only**

1. [9 Point] Review the course materials and answer the following questions
   1. To make real business impact with analytics, what are the 4Ps covered in this course?

The 4P’s covered are as follows:

**Problem**

It is defined as a concise description of the issue that needs to be addressed. “A problem well stated is half problem solved”. It is important to be as specific about the problem. It is expected to highlight the negative points of the current situation and its potential impact.

**Product**

A good or service that is used to fulfill requirements for solving the problem defined above.

**Process**

While implementing an analytical solution to a business problem, it is important to understand the process changes from business, operational and system standpoint.

**People**

People are key to change management. It is important to make sure that

human resource accepts and adapts to the change as quickly as possible to mark a

successful analytical product implementation.

* 1. Name the three important elements for an “analytics product”

Three important elements for an analytics product: -

**Data**

Data plays an important role in analytics product. Availability of data on time, dependency on other factors and its comprehensiveness should also be considered.

Data is expected to be thoroughly validated for accuracy and consistency and adaptable to change management.

**Methodology**

Methodology offers underlying theory for understanding which method can be applied to build an analytical product. It comprises of methods and principles that defines the means of data collection. A methodology is the process used for carrying out research and not a method itself. Constraints are the limitations on generalizability, applications to practice or utility of findings that are the result of the method chosen to build a good model

**Communication**

It is imperative to effectively communicate the finding of an analytics product. This involves knowing what the audience expects, developing a story line and focusing on delivering results as opposed to explaining the process. Communication should be more focus oriented.

* 1. What’s the most important thing to do before starting any analysis?

Data needs to think of data holistically before starting any analysis. Relevant variables need to be identified and examined thoroughly. It should also be ensured that these variables can be measured, collected and analyzed.

* 1. Suppose you are doing an analytical project with another part of your company for the first time. You have never used the data from that part of the company. List 5 important things related to data that you want to confirm before starting the project

1. Think of data holistically before starting any analysis
2. Differentiate data needed for analysis from data needed for decision since data needed for analysis is often just part of decision making criteria.
3. Determine whether variables can be measured, collected and analyzed
4. Need to consider both qualitative and quantitative data and whether available data is structured or unstructured.
5. If sufficient data is not available, find alternative methods of attaining data such as crowd sourcing.
   1. Name and briefly explain two valuable concepts related to applying descriptive analytics that you have learned in this course

Aggregation vs Granularity:

**Data Aggregation Vs Granularity:** Data aggregation is any process in which information is gathered and expressed in a summary form, for purposes such as statistical analysis. Whereas granularity refers to the level of detail of the data stored in a table. They are components of how to describe data. Before analyzing data using prescriptive techniques, the granularity of data should be determined. While it might seem easy to summarize and aggregate the data, it is wise to understand the level of granularity needed

**Relationships among objective metrics** (Current/ Historical data) and driver metrics define by Quantitative and qualitative data. Relationships also help to understand correlation and causation.

* 1. Suppose you have two different models that can score the likelihood to re-purchase among your existing customers. What criteria would you use to compare the quality between the two models?

R square value is the percentage of the response variable variation that is explained by a linear model. 100% indicates that the model explains all the variability of the response data around mean. We can use R Square for preliminary analysis. However, a high R square value may not always yield the desired result. Other factors such as Coefficients, independent variables and their P Value should be considered to gather more insights about both models.

* 1. Name a potential analytics technique for each of the following scenarios
     + Predicting performance levels without domain knowledge, some but limited historical performance data at somewhat aggregated level without performance driver information – Delphi method
     + Predicting performance levels, with limited domain knowledge, plenty of structured and unstructured data Machine Learning Method
     + Deciding appropriate actions for large number of customers who needs highly customized solutions in a highly stochastic environment - KNN

1. [9 Points] Answer the following questions as concisely as you can according to the scenarios described
   1. Suppose you have a friend who believes that Descriptive Analytics is just about the granularity, the speed, and the cost of doing so in collecting data. Now that you have taken “Managerial Analytics”, how would you explain to your friend (very concisely and convincingly) that his notion of Descriptive Analytics is not sufficient?

This notion of Descriptive Analytics is not sufficient as Descriptive Analytics focuses on following two aspects: -

1. What to describe
2. How to Describe

Granularity, speed and cost of doing so in collecting data is part of how to describe.

As part of What to describe, it is important to identify the objective metrics (historical data, depicting trends), key potential drivers and relationships among these two metrics.

* 1. You have been hired as a consultant to develop pricing model(s) for a CPG company. They have two major categories of products, sports drinks (6 major flavors, each of which has 3 sizes) and energy bars (8 different flavors, 3 sizes each, with diet and non-diet variants for reach). This company is only selling through its own stores. You have been given all the historical sales data from the past 5 years. What questions would you have up front? What data validation would you do accordingly? What other concerns would you have despite the data?

**Questions :**

Do we have past data for both the products?

What is the most popular product. What is the location of most popular product. Can we have different prices for different products.

Consumption of each product. For eg. Is diet in more demand than non-diet.

**Data Validation :**

How was this data collected. Is there any missing data in case a product was introduced later.

How has that been handled in analysis. Are there any outliers or unnecessary noise in the data.

**Other Concerns :**

Are there any other factors involved.

Do they have competitors in market with similar products.

Are the sales seasonal.

Do people purchase more if discounts/coupons are offered.

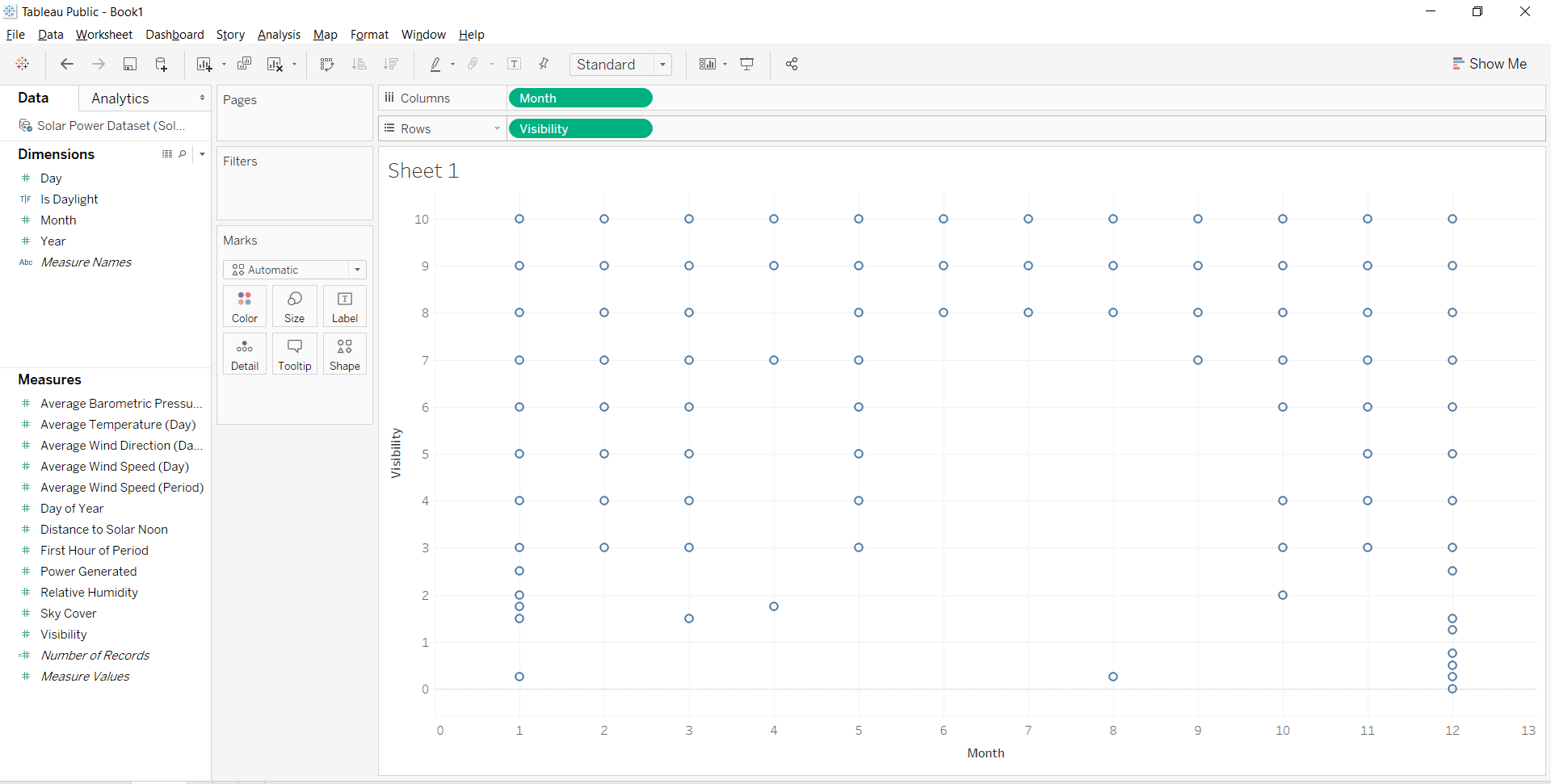
* 1. You are working for a movie company, and you are building a regression model to predict whether a given movie will be successful once it is launched. An analyst of yours indicates that the following variables have all provided incremental value in predicting movie performances. Would you include them, why or why not?
     + Gentry – **Include** as people from differing backgrounds/societies are usually selective what kind of movie they would be interested in.
     + Release month of the year – **Exclude,** release month would hold little importance as compared to release date/holiday/seasonal off etc.
     + How close the release was to Thanksgiving – **Include** Tendency of people going out is more during holiday time
     + Whether it is a sequel or not? – **Exclude**, A sequel movie’s performance might be more dependent on how well the original movie did. We cannot generalize if sequel movie would be good/bad.
     + Whether it is an animated film or not? – **Include** – Usually children are interested in animated films
     + Whether it is a foreign film or not? **Exclude**, we cannot generalize movie’s performance based on if it is a foreign film.
     + Whether it is a sci-fi film or not? **Include** A sci-fi is a genre of a movie that can be used to project if movie’s performance will be good or not.
     + Whether it is an action movie or not? **Include** popularity of action films can be used a s a measure to predict movie’s performance.
     + The number of past Oscar winners in the film? **Include** people tend to watch movies that have received acclaim in Oscars
     + The number of Oscar awards the director has won **Include –** Movie’s performance can be dependent on the credibility of the director.
     + Gender of the director **Exclude -** Gender of director should not decide the movie made by the director.
     + Whether the film has won some significant awards? **Includ**e When a movie has already won significant awards, its credibility has been built and it might do well
     + Movie critique rating **Include** People often follow movie’s reviews and critics ratings
     + The box office performance before when this is a sequel – **Include** Previous movie’s performance can be used to judge if people will come to watch the sequel.
     + The number of countries to be released at the same time – **Include** If released in more countries, it is likely to do well as more people will be able to see it.

Answer will be dependent on each variable individually. Descriptive Analytics should be used to gather past trends to understand if a movie will be successful or not.

1. [14 points] Download the “Solar Power Dataset” file from Canvas. This is a data file that can be used to predict the output of a solar power system.
   1. Are there any missing value or outliers? How would you like to treat them? Make sure that you apply appropriate treatment first.

Missing Values : Average speed wind(Period) for month 11 and Day 29 is missing. This is being removed from further analysis.

Yes, there are outliers. For month 8, value for visibility is considerably low for Day 27. It is probably a typo and should be 2.5 instead of 0.25.



These outliers and missing values will be removed from data for future analysis.

* 1. Are there any “noises” (i.e., inappropriate data) in the dataset that you would like to remove before modeling? If there are, remove them first.

Firstly, power generated is 0 for night time. This is a noise and can be removed.

Also , there are a few columns that are not required to build the predictive model.

-Day

-Day of the year

-First hour of period

-Average speed of wind(period)

* 1. Are there any pairs of variables that are suspects of collinearity? Demonstrate your reasoning and treatment.

We create a correlation matrix to determine if there is any collinearity among variables.



Considering correlation for variables with more than 0.4, following are the pairs :

Year and Day of Year

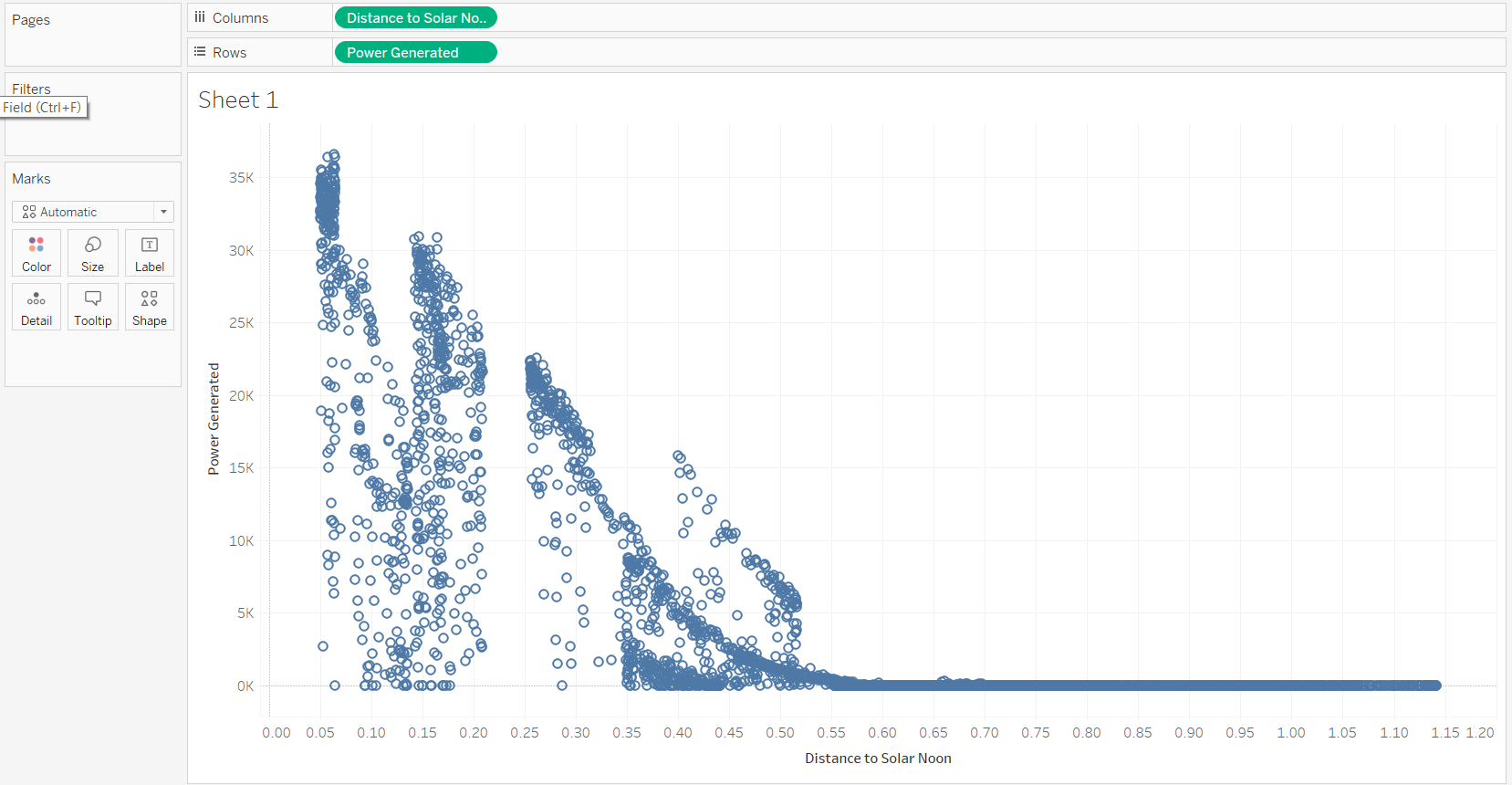
Year and year

Average wind speed(day) and average wind speed(period)

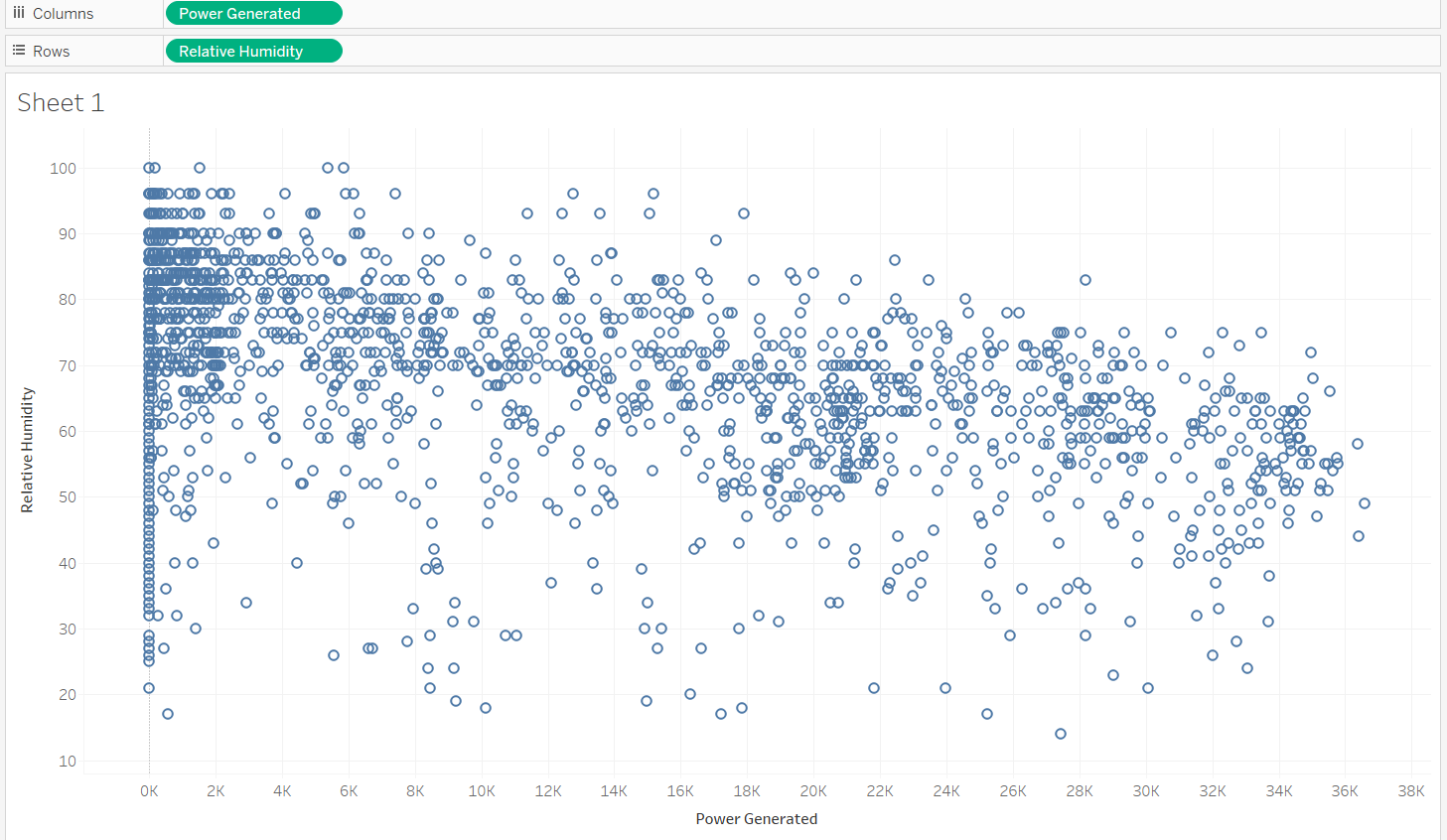
Theses variables should not be considered while building a predictive model.

* 1. Now you are about to build a predictive model to predict the power generated. Based on any preliminary analysis, which might be the first variable that you would include in your model, and which might be the second most important variable that you would include in your model?

Based on preliminary analysis, the most important variable that I would like to include is Distance to solar noon because of its high negative correlation with Power Generated as depicted in correlation matrix



Relative humidity is the second variable that we would like to include in model which has the second highest correlation -0.52 after Distance to Solar Moon.



* 1. Build a regression model in Excel using the above two variables to predict the power generated. Submit the model output. What’s the R-square?



R Square Value : 0.687

Power generated = 40063.4 -43127\*Distance to solar noon -221.9\*visibility

* 1. Would additional variables help such a predictive model? Show your eventual model and provide justification on why, or why not, additional variables help.

To improve the overall model, we add the following variables

-Sky Cover

-Average Wind Speed (Period)

-Average Temperature Delay



Improved R Value = 0.72

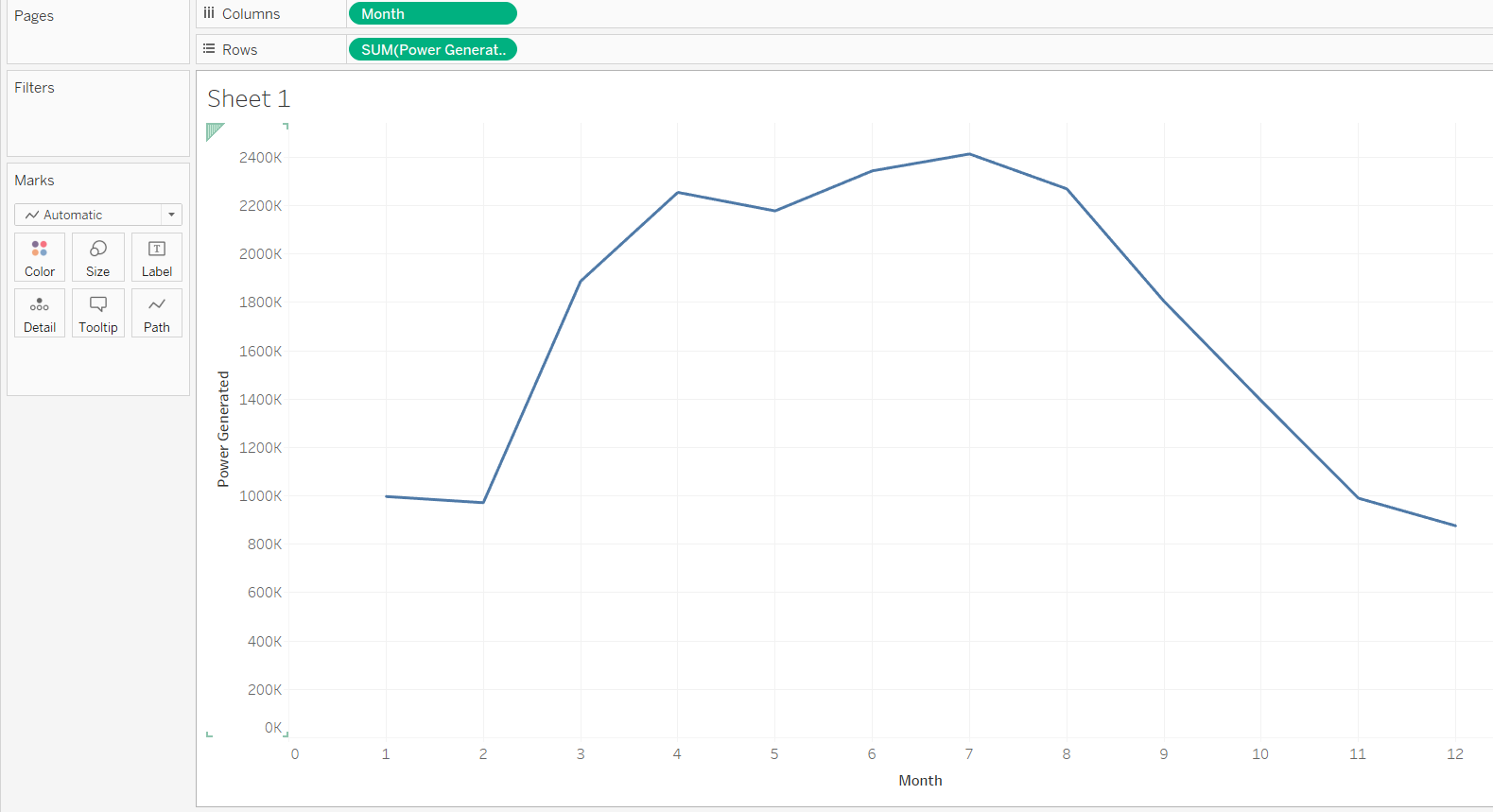
Adding more predictor variables increases the r-square value and hence provide a robust

model. Since all these variables affect power generation and explains the dependent

variable to some extent and hence when added to the model, increases the r-square value.

* 1. An analyst who wants to improve the model hypothesizes that more accurate model can be built if we divide the dataset by month and build a model separately for each of January, February… December. Apply your analytics mindset to determine whether this is a good idea or not.

When we plot month over Power generated, we observe that it varies significantly as shown below.



We create two different models for month 1 and 7.

For month 1



For month 7



Clearly Month 1 has lower R square value(0.50) and Month 7 has higher R Square value(0.81) so it might be a good idea to have separate models by month. The variables we are using (temperature, wind) all are expected to change with the course of time, so it makes sense to have different models for each month.

1. [10 points] Download the “March Madness” file from Canvas. This is the background data file for you to play in a “March Madness” pool. In this pool, your objective is to maximize the number of wins using the team that you have “purchased”. Each team has a cost. The higher the seed, the more expensive it gets. You will have a budget of $1.00. You are also provided with predicted probability of winning for each team at each stage. For example, Team E1 will have 97% probability of winning its first round. The subsequent probabilities are conditioned on the team’s ability to get there in the first place. For example, S12 has a 15% probability winning its game in the 4th round, if it can get there at the first place. In fact, due to potential matchups, S12 has only a 13% chance winning its 3rd round match. Since this is done based on the potential matchup, not all seeds at the same level have the same probability of winning at each stage. You will need to use Excel solver or any other optimization software fits you to solve the following.
2. Put together a team that will give you the highest “expected number of wins” without going over budget. Submit the list of teams. What’s the expected number of wins? (*Hint:  for conditional probability calculation:*

*Winning times for each team: X = {0, 1, 2, 3, 4, 5, 6}*

*P(X=0) = P(lose the first round)*

*P(X=1) = P(win the first round, lose the second round)*

*P(X=2) = P(win the first round, win the second round, lose the third round)*

*And so on.... )*

Using Integer programming, we find that 16 is expected number of wins required.

Selected teams are :

E4, E6, E12, S4,S5,S6,S7,S12,S13,M6,W1,W6,W13,W14

1. One potential problem with a) is that you might have teams that actually going to meet early. For example, you could have S3 and S14 both in your selected teams. Even though they may be part of the optimal answer, the fact that they are going to meet in the first round automatically eliminates one team from your portfolio. Try to set up the solution to prevent teams from meeting in the first round. Submit the list of teams. What’s the expected number of wins? (*Hint: In the first round, each team would compete within their own region, such as S1 vs S16, S2 vs S15, E1 vs E16 ... and so on.)*

Team chosen – E2, E4, E6, E12, S6, S7, S12, S13, M6, M13, W1, W6, W13, W14

Expected number of wins – 16

