

[← BACK TO GRADES](#)

Homework Assignment 2 (week 2)

Faculty Feedback

Score

Last published: 7/21/2019 7:09 PM PDT

100 / 100 (100.00%)

Comments

Comments

Week 2 Assignment Rubric and Explanations

Great start and please review the following to think about areas that can use improvement. All assignments can always be improved – which is the goal here. The most useful method is employ here is to compare your assignment to the notes below and to think about what is there, what was awesome, what was good, what is missing, and what was OK. Then think about ways to improve. This process of reviewing your own work is a great way to learn and to improve. It is also a very good method for learning to be critical and to perform analysis. Grade descriptions are below as I grade holistically. My notes are below as well.

NOTES:

BEAUTIFUL REPORT

FORMAT:

Thank you for following the format and including the code in your submission!!!

Overall -- Good work here.

INTRO:

Great introduction here. The problem, data, and goal are clearly defined and described. The table is well-set!

EDA:

Very good EDA, discussion and explanation. It is clear that you spent time poring over the data.

VIZ:

Great tables, plots and figures. They really support the discussion well and characterize the key points of the data!

CONCLUSIONS:

The conclusions are a bit slim, but expectedly so given the nature of the data. More will be expected next week.

Some General Notes Below:

Introduction: (2 - 3 paragraphs)

An Introduction is about the area or topic, not about the data or models. The introduction helps the reader to understand what the assignment area is about. For example, support the assignment is about schools. In this case, the introduction is about school systems, why schools are measured and ranked, who might be concerned with school measures and rankings (such as students, parents, states, governments, and funding agencies), and the value of comparing schools.

An introduction is a like a warm-up or like dating. It allows the reader to “get to know” the area of interest.

(1) If the Intro is less than 2 paragraphs (deduction -5). A paragraph is 7+ sentences.

(2) If the Intro is not written in clear and proper English, with correct grammar, sentence structure, and flow. (up to -10).

(3) If the Introduction has a “undergrad” feel, such as starting sentence with “My Assignment is about...”, or “I am going to talk about...”, or “In this Assignment, I will....”. At the graduate level, write Assignments as though they are technical papers. (up to -10 for non-graduate level writing).

(4) Avoid the use of “I”, “we”, “us”, “you”, ... Always remember that Assignments are about the area of interest, the topics, the models, the analysis, the results, and the outcomes/conclusions, they are not about you. (deduction: up to -10).

(5) The Intro should not contain any information about the dataset or the data cleaning, prep, processing, etc. Everything about the dataset goes into the Analysis section under the “About the Data” subsection. (up to -10 if dataset is discussed in the Intro).

Analysis and Models

The Analysis section contains subsections.

The first subsection is “**About the Data**” which contains all the information about the dataset, the variables, the cleaning and prep, checking for and dealing with missing values, checking for and dealing with incorrect values, checking for and dealing with outliers, feature generation, normalization (if needed), etc. In this subsection, you will also “explore” the data. This means that you write about each variable, visualize each variable, and talk about what the variable represents. Tables are great for this as well.

If you do not clean and prep the data, the deduction is up to -15). No matter how clean data “looks” - always write code to check it, update types, clean, and prep for analysis.

The second and remaining subsections of Analysis are the model(s).

In some cases, there may only be one model.

A model is any method used to analyze the data.

Each Assignment specifies which models to use.

Always include model details and parameter values when applicable.

*** Have Visualizations throughout the assignment.

Results

The Results section of the Assignment will have a subsection for results for each model (assuming that you have more than one).

Results are technical. They offer technical information about what was found in the analysis. For example, if you performed a correlation in the analysis between all pairs of numeric variables, then your results would discuss the r-value and relationship of each pair. Similarly, if you looked at measures of center and variation, the results talk about what those measures are and what they reveal. For example, if the mean is less than the median, the data is skewed, which means....

Each model we will use in this class has results and parameters associated with it. For example, association rule mining will offer the top ten rules for sup, conf, and/or lift if you code it to do so. These would go into the results along with the sup, conf, and lift for each rule. The meaning would also be discussed.

**** Always have visualizations**

Conclusions

2 - 3 paragraphs.

This area is not technical at all.

This area explains what was actually found in a way that would make sense to anyone. For example, if you discovered in the analysis that association rule mining with a conf of .2 and a sup of .3 offered 10 rules, you would talk about the measures and values and rules in the *results*. In the Conclusions, you would talk about what it all means. So you would not include the rules themselves or mention of technical measures such as conf or sup. Rather, you would say that you found (as a random example) that people who buy diapers are very likely to buy beer and that this means that a store should consider placing these items “near” to each other.

Basis For Grades:

100: This means that your Assignment was amazing and so perfect that nothing can be improved. It covered everything – cleaning, prep – analysis that makes sense – visualizations – results (that are true) – etc. There is nothing really left to improve.

95: This means that your Assignment is really good! You covered most of the items noted below and perhaps a few others not noted. You can make some improvement on pre-processing and results analysis, as well as perhaps other visualizations. Overall – you have the idea and you did well.

90: This means that your Assignment is good, but could be a little better. Perhaps add items such as further data cleaning and pre-processing, data normalization, better or more visualizations, and/or more robust conclusions. Many students forgot to change Section to a

factor for example. Very few students summed and normalized the data to look at the percentages for each attribute.

85: This means that your Assignment is a good start and largely meets the more general and overall requirements. Here, you used R, you did some analysis, you did some cleaning, you made some graphs, and you reached some conclusions. However, there is room for improvement.

Below 85 means that the level of 85 above was not quite met and many elements were missing.

See my notes below for details about elements. This is not an exhaustive list – but will help.

*** As we move forward – the grading levels will become more strict and expectations will increase.

*** If you are accidentally missing (or did not submit) the .R, I will not grade off (deduct points) for this Assignment – but for future assignments – be sure to include the .R file so that I can copy it and run it. Make sure you open the file directly and not from a path.

NOTES:

The data used for this assignment is the storyteller data.

1) I recommend saving the data as .csv and updating all variable names so that they have no spaces.

2) Here is **a copy** of my updated

data: https://drive.google.com/file/d/1s7bVRKAwiSKis3bpys0X_gdtMt6_FqoA/view?usp=sharing

Assignment Description:

1) Look at and think about the Data first. What does it seem to mean? What are the attributes (variables)? Here we have 5 schools (A – E) that are all implementing (teaching) the same math class. The Math Class has 35 lessons. At the time of the data collection, the term was 75% complete. This means that to be right on target (assuming equal lesson duration), all students should be on Lesson $.75 \times 35 = 26$ (rounded down). So, a student can be above or behind in the class. There are only 35 lessons. The data looks at how many schools are

a. VeryAhead (+5),

- b. Middle (0 to 5 ahead),
- c. Behind (1 to 5 below),
- d. More Behind (6 to 10 below),
- e. Very Behind (more than 10 below), or
- f. Completed (finished the course).

3) What is the *story* that this data tells?

4) Use R, analysis, and visualization.

Important Elements:

Assignments are like academic or professional papers. This means they should be written using graduate-level technical writing. They should have proper headings: Introduction (explain what the topic is, its value, overall goals, etc).

More hints and help for Assignment 2....

1) The Week 2 Assignment offered you a dataset and asked you to tell a true story about the data.

2) This is never easy and there is often more than one viable story – depending on focus and perspective.

3) For any dataset, the first step is to clean the data. At the very least – this should include looking at:

- a. Missing values
- b. Correct values with incorrect representations (such as 0 instead of male)
- c. Incorrect values (such as .33 for age)
- d. Duplicates – depending on the nature of the dataset.
- e. Beyond the above – you can also look at outliers – though this did not apply in this case.
- f. ** Always write code to do the above, even if you do not think it is necessary. You never know what you will find – if you look.

4) The next step includes looking at each data type and correcting the errors. For example, Section is not an int and must be changed to factor (as it is qual and nominal).

5) Think about normalizing if you plan to compare. In this case – you must compare Schools. To do this, you must sum up (aggregate) the data by School and normalize each School by dividing each of its variable values by the total number of students. If you do not normalize – you may reach incorrect results. For example, School B actually has the highest percentage of students Middling. Consider the idea that School A has more students, and so it is likely that they will have more (in count) above and more (in count) below. So you must normalize by the total.

6) Consider which variables in the dataset you will be able to use. Notice that while “Section” is a variable, some of the Schools have only one section. Therefore, you will not be able to compare between school for this attribute. Next, the only question you can really ask is if sections seem significantly different within one school. However, this would require an ANOVA test to confirm.

7) If you plan to use correlation – make sure that it makes sense. For example, if you want to look to see if there is a linear relationship between Middling and Behind1-5, first ask yourself what this would tell you. Suppose there is a correlation. What does this mean? Does it mean anything? Because only Schools A and B have enough sections, you can only see if there are correlation for these two and between these two. However, you may find little information there.

8) The next steps included an iterative (back and forth) process where you think about what you are trying to investigate about the data and write code to investigate it. Based on findings, you will generate directions for investigation and other questions.

9) For your Assignments – especially those like this one – offer an overview of the initial thought, your thought process, your findings, your exploration, and then your results and conclusions.

10) Include visualizations and all R code that you use.

11) For visualizations – MAKE SURE THEY ARE LABELED :) Titles, clear axes, etc. If they are not readable, they are not helpful.

12) **** IMPORTANT **** **Always submit R code that runs.**

FOR ALL ASSIGNMENTS - YOU WILL SUBMIT the Assignment as .docx and the R code as .R.

13) For all future assignments – PLEASE SUBMIT your R CODE as a separate file .R. If your R code opens a file – submit code such that the file is in the SAME LOCATION as your code. In other words, open the file as “filename” rather than “C:/Users/R/whatever/whatever/....”. Remember, your paths are not my paths :)

14) **At the end – always have a paragraph that offers a conclusion.** For example, ...overall School B shows the best overall performance While School C is.... School E offersetc. etc.

VIZ:

Great tables, plots and figures. They really support the discussion well and characterize the key points of the data!

You may wish to introduce more and varied VIZ to help discuss and support your EDA, such as bar charts, histograms, tables, etc ...

The plots and visualizations are also great but few in number. Some more visual exploration of the data may shed some more insights; moreover the descriptions and explanations of the plots are somewhat topical, lacking a bit of detail. Be sure to explain your plots and discuss what one can infer from them.

CONCLUSIONS:

The conclusions clearly follow from the EDA and discussion. They are well formed and well thought-out. Good notes about more data – more is better!

The conclusions are also a bit slim and could use more bulk. Otherwise -- great work and I look forward to your Assignment #3 submission!

FORMAT:

Thank you for following the format and including the code in your submission!!!

Please include the code with your submission as appropriate.

Also -- please adhere to the standard report format.

Some General Notes Below:

Introduction: (2 - 3 paragraphs)

An Introduction is about the area or topic, not about the data or models. The introduction helps the reader to understand what the assignment area is about. For example, support the assignment is about schools. In this case, the introduction is about school systems, why

schools are measured and ranked, who might be concerned with school measures and rankings (such as students, parents, states, governments, and funding agencies), and the value of comparing schools.

An introduction is a like a warm-up or like dating. It allows the reader to “get to know” the area of interest.

(1) If the Intro is less than 2 paragraphs (deduction -5). A paragraph is 7+ sentences.

(2) If the Intro is not written in clear and proper English, with correct grammar, sentence structure, and flow. (up to -10).

(3) If the Introduction has a “undergrad” feel, such as starting sentence with “My Assignment is about...”, or “I am going to talk about...”, or “In this Assignment, I will....”. At the graduate level, write Assignments as though they are technical papers. (up to -10 for non-graduate level writing).

(4) Avoid the use of “I”, “we”, “us”, “you”, ... Always remember that Assignments are about the area of interest, the topics, the models, the analysis, the results, and the outcomes/conclusions, they are not about you. (deduction: up to -10).

(5) The Intro should not contain any information about the dataset or the data cleaning, prep, processing, etc. Everything about the dataset goes into the Analysis section under the “About the Data” subsection. (up to -10 if dataset is discussed in the Intro).

Analysis and Models

The Analysis section contains subsections.

The first subsection is “**About the Data**” which contains all the information about the dataset, the variables, the cleaning and prep, checking for an dealing with missing values, checking for and dealing with incorrect values, checking for an dealing with outliers, feature generation, normalization (if needed), etc. In this subsection, you will also “explore” the data. This means that you write about each variable, visualize each variable, and talk about what the variable represents. Tables are great for this as well.

If you do not clean and prep the data, the deduction is up to -15). No matter how clean data “looks” - always write code to check it, update types, clean, and prep for analysis.

The second and remaining subsections of Analysis are the model(s).

In some cases, there may only be one model.

A model is any method used to analyze the data.

Each Assignment specifies which models to use.

Always include model details and parameter values when applicable.

*** Have Visualizations throughout the assignment.

Results

The Results section of the Assignment will have a subsection for results for each model (assuming that you have more than one).

Results are technical. They offer technical information about what was found in the analysis. For example, if you performed a correlation in the analysis between all pairs of numeric variables, then your results would discuss the r-value and relationship of each pair. Similarly, if you looked at measures of center and variation, the results talk about what those measures are and what they reveal. For example, if the mean is less than the median, the data is skewed, which means....

Each model we will use in this class has results and parameters associated with it. For example, association rule mining will offer the top ten rules for sup, conf, and/or lift if you code it to do so. These would go into the results along with the sup, conf, and lift for each rule. The meaning would also be discussed.

** Always have visualizations

Conclusions

2 - 3 paragraphs.

This area is not technical at all.

This area explains what was actually found in a way that would make sense to anyone. For example, if you discovered in the analysis that association rule mining with a conf of .2 and a sup of .3 offered 10 rules, you would talk about the measures and values and rules in the *results*. In the Conclusions, you would talk about what it all means. So you would not include the rules themselves or mention of technical measures such as conf or sup. Rather, you would say that you found (as a random example) that people who buy diapers are very likely to buy beer and that this means that a store should consider placing these items "near" to each other.

Basis For Grades:

100: This means that your Assignment was amazing and so perfect that nothing can be improved. It covered everything – cleaning, prep – analysis that makes sense –

visualizations – results (that are true) – etc. There is nothing really left to improve.

95: This means that your Assignment is really good! You covered most of the items noted below and perhaps a few others not noted. You can make some improvement on pre-processing and results analysis, as well as perhaps other visualizations. Overall – you have the idea and you did well.

90: This means that your Assignment is good, but could be a little better. Perhaps add items such as further data cleaning and pre-processing, data normalization, better or more visualizations, and/or more robust conclusions. Many students forgot to change Section to a factor for example. Very few students summed and normalized the data to look at the percentages for each attribute.

85: This means that your Assignment is a good start and largely meets the more general and overall requirements. Here, you used R, you did some analysis, you did some cleaning, you made some graphs, and you reached some conclusions. However, there is room for improvement.

Below 85 means that the level of 85 above was not quite met and many elements were missing.

See my notes below for details about elements. This is not an exhaustive list – but will help.

*** As we move forward – the grading levels will become more strict and expectations will increase.

*** If you are accidentally missing (or did not submit) the .R, I will not grade off (deduct points) for this Assignment – but for future assignments – be sure to include the .R file so that I can copy it and run it. Make sure you open the file directly and not from a path.

NOTES:

The data used for this assignment is the storyteller data.

1) I recommend saving the data as .csv and updating all variable names so that they have no spaces.

2) Here is **a copy** of my updated

data: https://drive.google.com/file/d/1s7bVRKAwiSKis3bpys0X_gdtMt6_FqoA/view?usp=sharing

Assignment Description:

1) Look at and think about the Data first. What does it seem to mean? What are the attributes (variables)? Here we have 5 schools (A – E) that are all implementing (teaching) the same math class. The Math Class has 35 lessons. At the time of the data collection, the term was 75% complete. This means that to be right on target (assuming equal lesson duration), all students should be on Lesson $.75 \times 35 = 26$ (rounded down). So, a student can be above or behind in the class. There are only 35 lessons. The data looks at how many schools are

- a. VeryAhead (+5),
- b. Middle (0 to 5 ahead),
- c. Behind (1 to 5 below),
- d. More Behind (6 to 10 below),
- e. Very Behind (more than 10 below), or
- f. Completed (finished the course).

3) What is the *story* that this data tells?

4) Use R, analysis, and visualization.

Important Elements:

Assignments are like academic or professional papers. This means they should be written using graduate-level technical writing. They should have proper headings: Introduction (explain what the topic is, its value, overall goals, etc).

More hints and help for Assignment 2....

1) The Week 2 Assignment offered you a dataset and asked you to tell a true story about the data.

2) This is never easy and there is often more than one viable story – depending on focus and perspective.

3) For any dataset, the first step is to clean the data. At the very least – this should include looking at:

- a. Missing values
- b. Correct values with incorrect representations (such as 0 instead of male)
- c. Incorrect values (such as .33 for age)

d. Duplicates – depending on the nature of the dataset.

e. Beyond the above – you can also look at outliers – though this did not apply in this case.

f. ** Always write code to do the above, even if you do not think it is necessary. You never know what you will find – if you look.

4) The next step includes looking at each data type and correcting the errors. For example, Section is not an int and must be changed to factor (as it is qual and nominal).

5) Think about normalizing if you plan to compare. In this case – you must compare Schools. To do this, you must sum up (aggregate) the data by School and normalize each School by dividing each of its variable values by the total number of students. If you do not normalize – you may reach incorrect results. For example, School B actually has the highest percentage of students Middling. Consider the idea that School A has more students, and so it is likely that they will have more (in count) above and more (in count) below. So you must normalize by the total.

6) Consider which variables in the dataset you will be able to use. Notice that while “Section” is a variable, some of the Schools have only one section. Therefore, you will not be able to compare between school for this attribute. Next, the only question you can really ask is if sections seem significantly different within one school. However, this would require an ANOVA test to confirm.

7) If you plan to use correlation – make sure that it makes sense. For example, if you want to look to see if there is a linear relationship between Middling and Behind1-5, first ask yourself what this would tell you. Suppose there is a correlation. What does this mean? Does it mean anything? Because only Schools A and B have enough sections, you can only see if there are correlation for these two and between these two. However, you may find little information there.

8) The next steps included an iterative (back and forth) process where you think about what you are trying to investigate about the data and write code to investigate it. Based on findings, you will generate directions for investigation and other questions.

9) For your Assignments – especially those like this one – offer an overview of the initial thought, your thought process, your findings, your exploration, and then your results and conclusions.

10) Include visualizations and all R code that you use.

11) For visualizations – MAKE SURE THEY ARE LABELED :) Titles, clear axes, etc. If they are not readable, they are not helpful.

12) **** IMPORTANT **** **Always submit R code that runs.**

FOR ALL ASSIGNMENTS - YOU WILL SUBMIT the Assignment as .docx and the R code as .R.

13) For all future assignments – PLEASE SUBMIT your R CODE as a separate file .R. If your R code opens a file – submit code such that the file is in the SAME LOCATION as your code. In other words, open the file as “filename” rather than “C:/Users/R/whatever/whatever/....”. Remember, your paths are not my paths :)

14) **At the end – always have a paragraph that offers a conclusion.** For example, ...overall School B shows the best overall performance While School C is.... School E offersetc. etc.


Student Submission | [Homework Assignment 2 \(week 2\)](#)

Response

Last submitted: 7/17/2019 12:38 PM PDT

No response

Files | [Download all \(2\)](#)

File Name	Uploaded	Feedback
Ryan_Timbrook_HW2.doc	7/17/2019 12:37 PM PDT	
ist707_week2_hw.R	7/17/2019 12:37 PM PDT	