Activity: Intro to Data Mining Tools | Trees, Classification Rules, and Association Rules

ⓘ This is a preview of the published version of the quiz

Started: May 21 at 8:11pm

Quiz Instructions

Intro to Data Mining Tools

Trees, Classification Rules, and Association Rules

We reviewed both the Orange and WEKA tools. Now it's time to get some hands-on experience!

Orange:

*"Orange (https://orange.biolab.si/) is a data mining and visualization toolbox for novice and expert alike. To explore data with Orange, one requires **no** programming or in-depth mathematical knowledge. We believe that workflow-based data science tools democratize data science by hiding complex underlying mechanics and exposing intuitive concepts. Anyone who owns data, or is motivated to peek into data, should have the means to do so." -* **Orange Github (https://github.com/biolab/orange3)**

Weka:

"Weka is tried and tested open-source machine learning software that can be accessed through a graphical user interface, standard terminal applications, or a Java API. It is widely used for teaching, research,

and industrial applications, contains a plethora of built-in tools for standard machine learning tasks, and additionally gives transparent access to well-known toolboxes such as ***scikit-learn*** *(https://markahall.blogspot.co.nz/2015/06/cpython-integration-in-weka.html) ,* ***R*** *(https://markahall.blogspot.com/2012/07/r-integration-in-weka.html) , and* ***Deeplearning4j*** *(https://deeplearning.cms.waikato.ac.nz) ." -* **Weka** **(https://www.cs.waikato.ac.nz/ml/weka/)**

The Data Science Pipeline

Through this activity, we will also get experience with the overall **data science pipeline.** Through this "pipeline," we are entering data with the goal of obtaining insights. The steps of this pipeline include:

1. Obtaining our data
2. Scrubbing or cleaning our data (or "data preprocessing")
3. Exploring, visualizing, and gaining a better understanding of our data (e.g., finding patterns)
4. Modeling our data (e.g., creating decision trees or finding classification rules)
5. Interpreting our data

Additional resource: **https://towardsdatascience.com/a-beginners-guide-to-the-data-science-pipeline-a4904b2d8ad3 (https://towardsdatascience.com/a-beginners-guide-to-the-data-science-pipeline-a4904b2d8ad3)**

Step 1

Download and install the tools

You've seen a short demonstration of both tools. For this activity, we will be working to create some decision trees, classification rules, and association rules. You can select one of the tools to focus on or try exploring both. For the first step, we need to **download and install the tool(s)** we will choose to work with.

Orange:

Go to *https://orangedatamining.com/* and click *Downloads*.

Weka: ***(https://waikato.github.io/weka-wiki/downloading_weka/)***

Download at **https://waikato.github.io/weka-wiki/downloading_weka/ (https://waikato.github.io/weka-wiki/downloading_weka/)**

Question 1 **0 pts**

Step 2

Choose your data

Now, let's choose the data we want to work with. Some of these I found based off of your interests (from the survey given last week), and some I included because they looked like good datasets to begin with.

Health: **https://www.kaggle.com/nareshbhat/health-care-data-set-**

on-heart-attack-possibility (https://www.kaggle.com/nareshbhat/health-care-data-set-on-heart-attack-possibility)

HR Analytics: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset (https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset)

Banking: https://www.kaggle.com/kidoen/bank-customers-data (https://www.kaggle.com/kidoen/bank-customers-data)

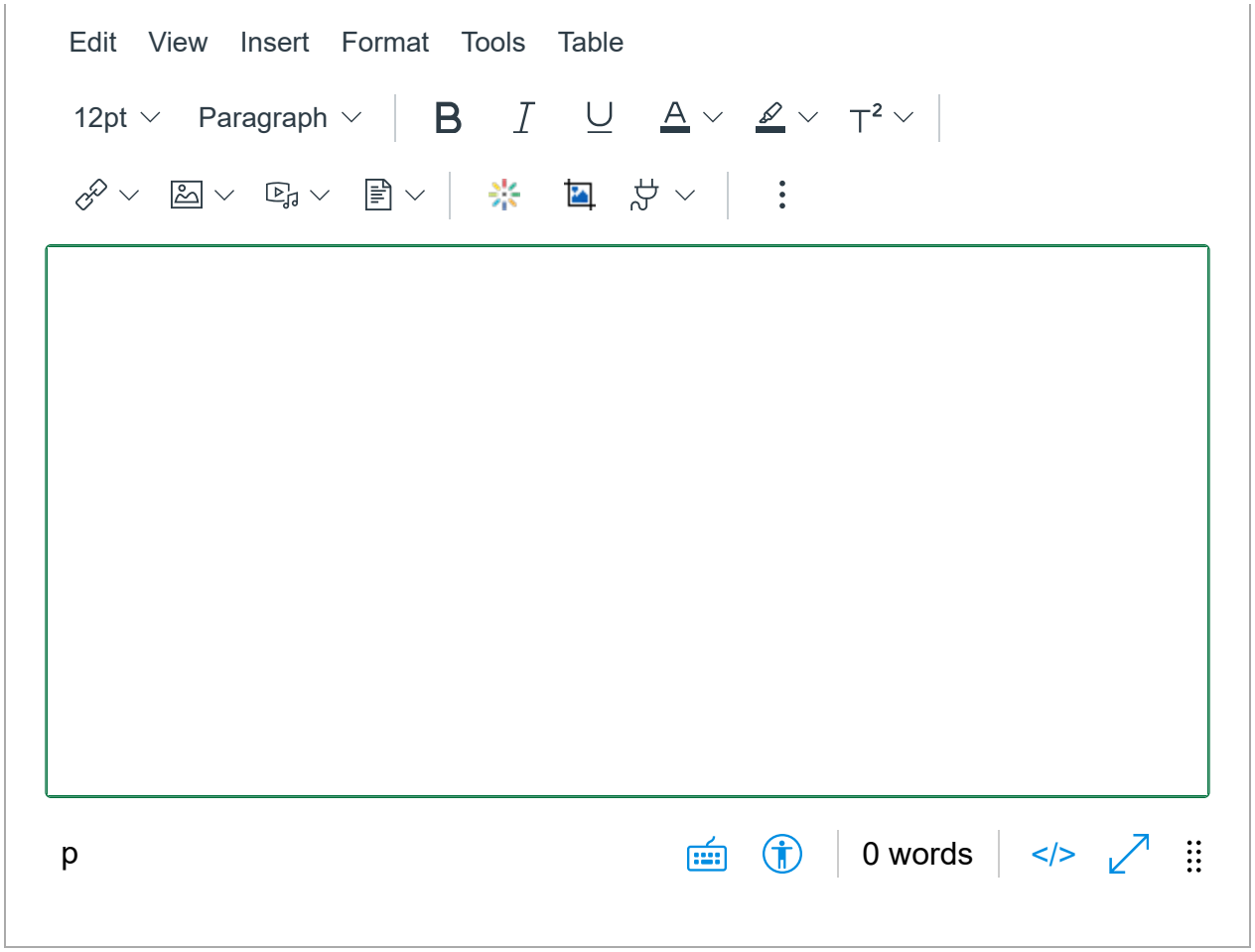
Wine Quality: https://www.kaggle.com/indranisen06/wine-quality-dataset (https://www.kaggle.com/indranisen06/wine-quality-dataset)

Politics: https://www.kaggle.com/johnwdata/2016-election-county-election-data (https://www.kaggle.com/johnwdata/2016-election-county-election-data)

Once you select your dataset, answer the following:

What would be the problem you are trying to solve with this dataset? For example, in our mushroom dataset example, we were trying to classify which mushrooms were edible or poisonous.

What are the features you would be exploring in order to solve your problem?



Instructions for Orange

Once you have selected your data, start a new file in Orange. To **load in your data**, drag in a "File" widget (located under the "Data" category). Then, you can double click on the File widget to select which file you want to load in. Once it is loaded, you can then close out.

Instructions for Weka

After opening Weka, select the "Explore" option. In here, you can then click "Open file..." to choose your data file to load in. Make sure to also change the "Files of Type" to "*.csv" (or whatever your file type is). Your data should now be loaded, and you can immediately see some information such as your attributes and counts per attribute type.

Step 3

Clean your data

"*Garbage in, garbage out.*"

It is important to check the quality of the data before proceeding. If we have bad data (inputs), then we'll have bad outputs (e.g., our models will be inaccurate).

In data cleaning, methods are usually applied to either remove, correct, fix incorrect data. For example:

- **Irrelevant data -** This is data that isn't entirely needed for the context of the problem we are solving. If we are absolutely sure that a piece of data is irrelevant, it may be dropped. Otherwise, we may keep most features to explore the relationships between variables (e.g., which features are more important for the problem).
- **Duplicates -** These are data points that are repeated. Some examples include a user being entered into a dataset twice by mistake, or an article/piece of text being scraped more than once. These duplicates can be removed.
- **Missing or null values -** If we have missing data, there are multiple ways we can choose to deal with them. **(1) Drop.** If the missing values are scarce, we may be able to simply drop those rows. If a given column has many missing values, we may choose to drop the whole column. **(2) Impute.** There are different methods for calculating the missing value. For example, one can use statistical values (e.g., mean or median) or random values to replace the missing value. **(3) Flag.** A third option is to "flag" those missing values, since saying the data is missing can also be informative.

For example, missing data could be filled out with 0's (numerical) or "Missing" (categorical). One thing to note though is that when filling out with 0's (for example), it is important that those zeroes are not being used in any statistical calculations.

Additional Resources:

- **(https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4)** https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4
- **(https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4) https://en.wikipedia.org/wiki/Data_cleansing#Data_quality (https://en.wikipedia.org/wiki/Data_cleansing#Data_quality)**

Question 2 **0 pts**

For this activity, let's just focus for now on dealing with any **missing values.**

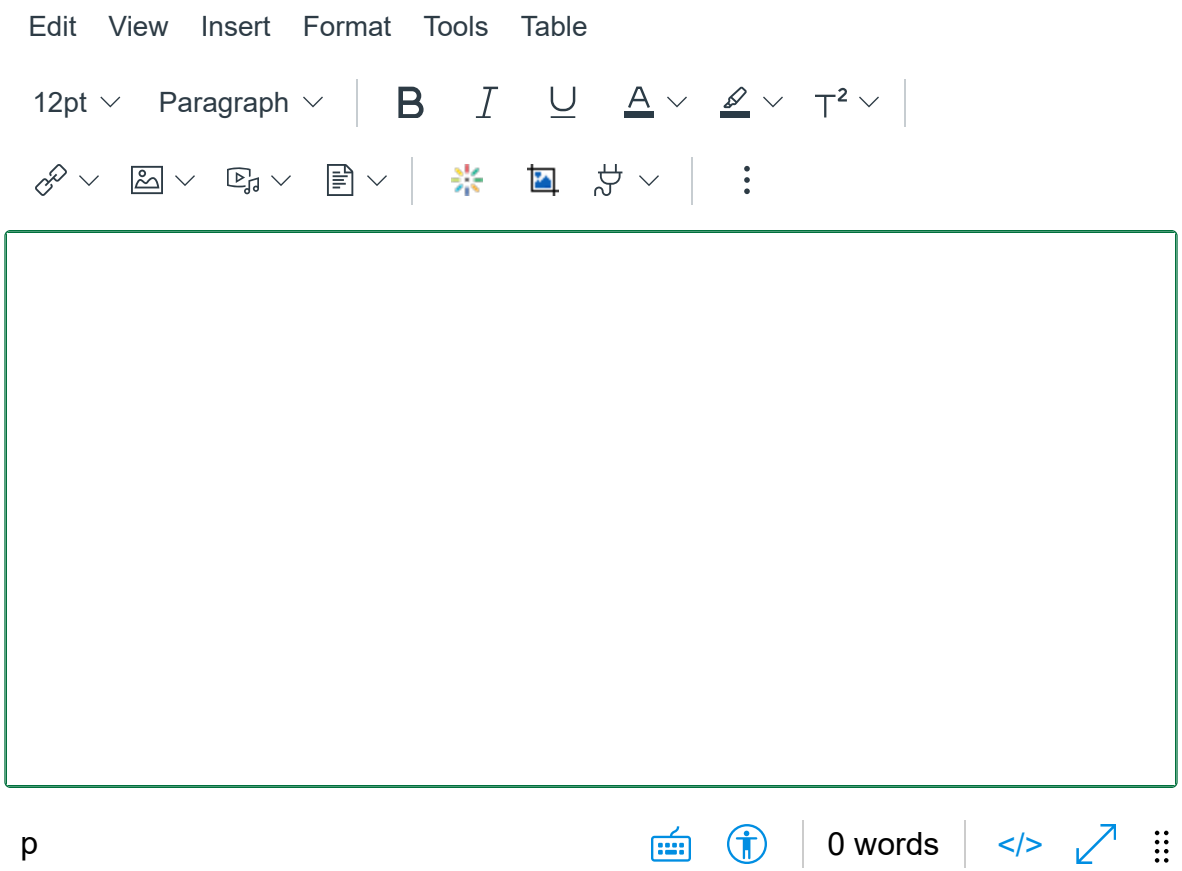
Take a look at your data. Do you see missing values? How will you deal with those missing values (perhaps the simplest way just to begin would be to remove them)?

Hints:

Orange - You can use the "Data Table" widget to view your data. The "Feature Statistics" widget also has a lot of helpful information (like a missing values count). Then, take a look at the "Preprocess" widget to remove sparse features and/or impute missing values (if needed).

Weka - In the first window after loading the data, you can select attributes (in the "Attributes" section) to view various statistics (in the "Selected attribute" section) such as the counts of missing values. You

can then select which attributes to remove (back in the "Attributes" section). You can also click the "Edit..." button at the top to manually go through the data to edit or delete rows.



Step 4

Understand your data

We will be beginning with a **classification problem** in order to get some practice with generating decision trees and classification rules.

Classification is the task of assigning labels to unlabeled data instances. A **classification model**, then, represents the relationship

between the attribute set (features) and the class labels. A **decision tree** is a type of classification model. And once that model is created, we can then use it to find our classification rules.

But before we dive into classification, we need to understand our data and what patterns may exist.

Start by thinking about the following questions.

Question 3 — 0 pts

With your chosen dataset, what are you trying to classify? What is the **target variable** (or class label)?

Edit View Insert Format Tools Table

12pt Paragraph

p 0 words

Question 4 — 0 pts

What **attributes** (or features) do you plan to use for classification? Do any of the attributes seem irrelevant (that can then be ignored)? Do any of the features seem like they will be more important?

Edit View Insert Format Tools Table

12pt Paragraph B I U A T²

p 0 words </>

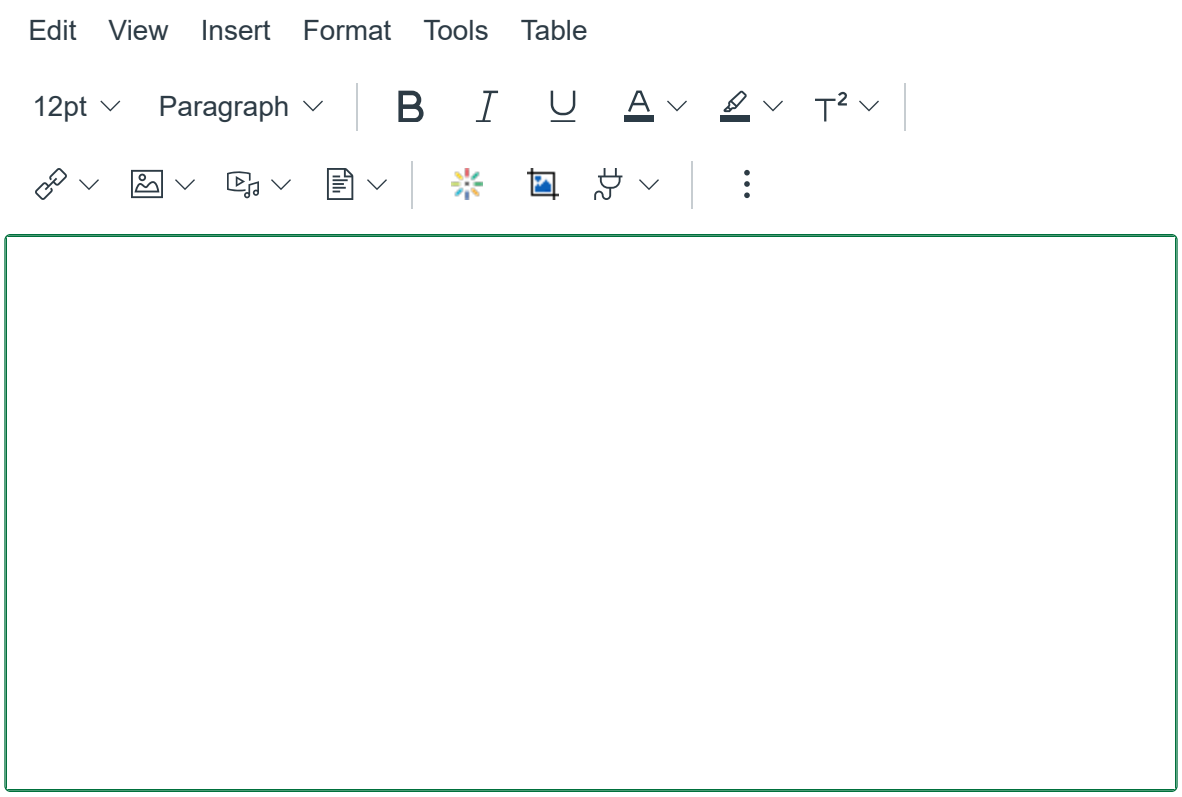
Question 5 — 0 pts

Do you see any relationships between your target variable and attributes?

Hint: Try exploring various methods of visualization.

Orange - Look at options under the "Visualize" tab.

Weka - Under the "Preprocess" tab in the bottom right screen (displaying distributions), you can select which column to display. You can also select the class (your target variable) and click "Visualize All." You may also take a look at the "Visualize" tab (tabs at the top).



Step 5

Model your data: decision trees

Now that you have decided on the **attributes** and the **target variable**, let's create a **decision tree**.

Resources on how to do this are provided below for each tool:

Orange:

https://orangedatamining.com/workflows/Classification/ (https://orangedatamining.com/workflows/Classification/)
Scroll down to the "Classification Tree". This is a base workflow for the decision (or classification) tree. Based on your data, you may need to include additional widgets for selecting the columns (features and class label).

Weka: Go to the "Classify" tab. Under the "Classifier" option, click "Choose." Go to classifiers > trees > J48. Under "Test options" you may select "Cross-validation" or "Percentage split." For the dropdown right below Test options, select your **target variable.** Then click "Start."

Step 6

Interpret your decision tree

Now that the decision tree has been created, you need to be able to view it.

Orange - This is done using the "Tree Viewer" widget (shown in the overall workflow given in the previous step).

Weka - Under the "Result list" section, right click the tree results you wish to visualize and then click "Visualize tree."

Question 6 **0 pts**

Before we move on to generating classification rules, take a look at the tree you have just created. What are some classification rules you can glean just from taking a look at your tree?

Edit View Insert Format Tools Table

12pt Paragraph

p 0 words

Step 7

Model your data: classification rules

Now that we have created and viewed our decision tree, we will move on to generating classification rules.

Orange - For this one, use the "CN2 Rule Induction" widget.

Weka - Once again, select the "Choose" button under "Classifier." Then select classifiers > rules > RoughSet. Then, select your "Test options", your target variable, and click Start.*

* Note that for the classification rules in Weka, you can generate the rules and see the overall accuracies (how well the rules describe the dataset), but you are unable to view the actual rules like in Orange.

Question 7 **0 pts**

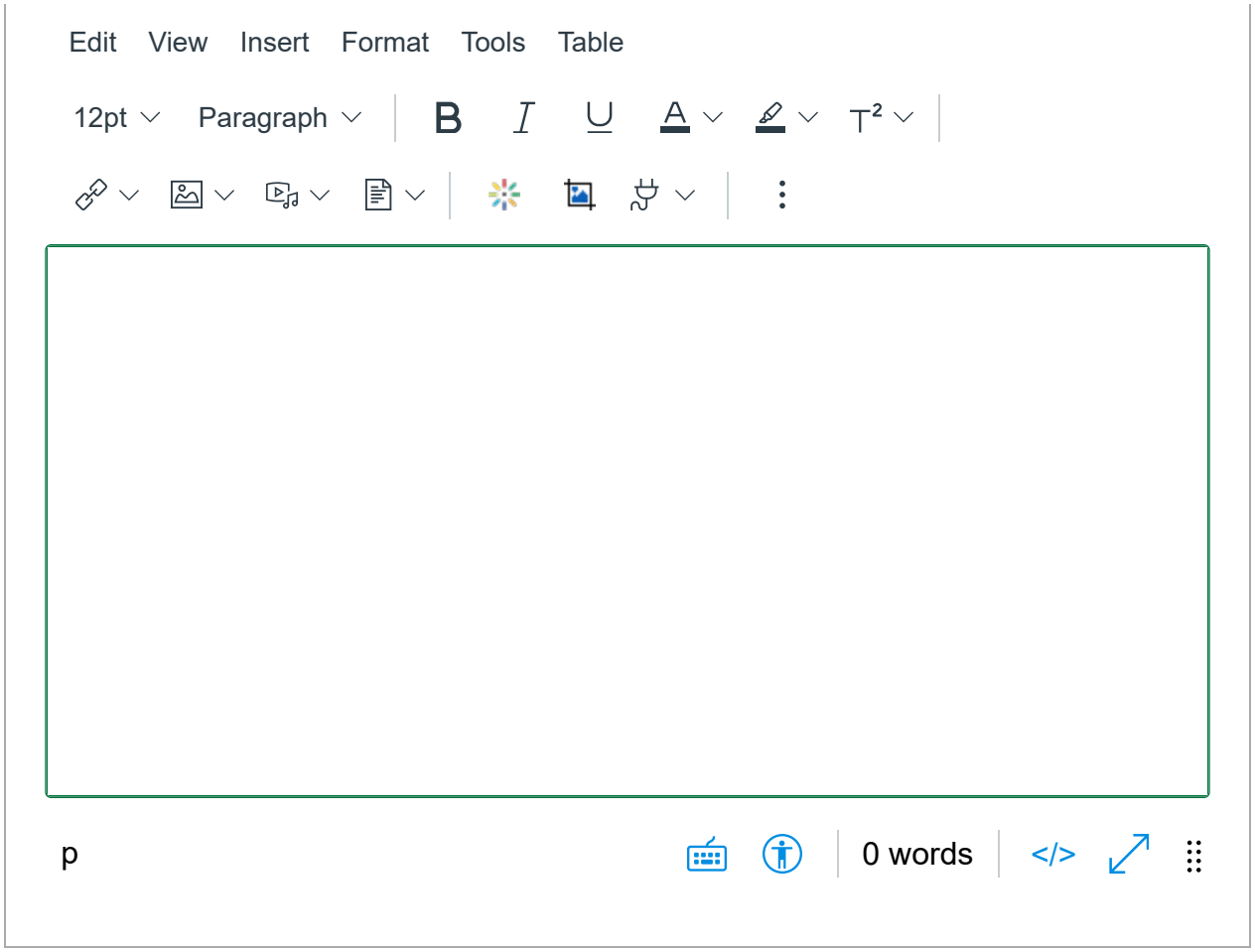
Step 8

Interpret your classification rules

Let's take a look at the rules now.

Orange - Use the "CN2 Rule Viewer" widget.

Discuss some of the rules you found and provide examples. Are there any that stand out to you?



Step 9

Model your data: association rules

Finally, we will get some experience with association rules.

Orange - In order to generate association rules in Orange, you will need to install additional add-ons. From the toolbar, select Options > Add-ons. Then select "Orange3-Associate." Once you have obtained this, there should be an additional section available called "Associate." Use the "Association Rules" widget, select your "Minimal support", "Minimal confidence", and "Max. number of rules." Then click "Find Rules."

Weka - Select the "Associate" tab. We can leave the "Associator" as "Apriori." Then click start.

Question 8 — 0 pts

Step 10

Interpret your association rules

Let's take a look at the rules now. Give some examples of association rules you found and discuss them briefly. For example, did they make sense to you? Why or why not? Did anything stand out?

Edit View Insert Format Tools Table

12pt Paragraph

p 0 words

Question 9 **0 pts**

Step 11

Reflect on your learning

Briefly reflect on what you learned today (the overall data science pipeline, getting some hands-on experience with the tool(s), etc.). Think about questions like how this could help you with your future goals, anything specific that was interesting (any "aha!" moments?), or perhaps fuzzy areas you may need to review again.

Reflection is an important part of learning as it is the step that helps us bring together our experiences to make sense of and grow from it.

Also, please let me know if there were any parts of this activity that were confusing or could be improved upon! :)

Edit View Insert Format Tools Table

12pt Paragraph B I U A T²

p 0 words </>

Not saved Submit Quiz