**Question 6**

|  |  |  |
| --- | --- | --- |
| Question | Assume the Pandas DataFrame below is assigned to the variable df. Furthermore, assume that df[“fare”] is in dollar units.  Which of the following lines of code will convert df[‘fare’] into euro units? At the time of this writing dollar:euro ratio is 0.92 (1 dollar = 0.92 euro). Mark each line as True or False. | |
| Image |  | |
| Option 1 | df[“fare”] = df[“fare”].apply(lambda x: x\*0.92) | **True** |
| Option 2 | df[“fare”] = df[“fare”].mul(0.92) | **True** |
| Option 3 | df[“fare”] = df[“fare”].multiply(0.92) | **True** |
| Option 4 | df[“fare”] = df[“fare”].transform(lambda x: x\*0.92) | **True** |
| Option 5 | df[“fare”].update(df[“fare”] \* 0.92) | **True** |
| Option 6 | df[“fare”] = df[“fare”] \* 0.92 | **True** |
| Correct Feedback: | Correct! These are all ways to perform this conversion on a Pandas Series. | |
| Incorrect Feedback: | Incorrect – please try again! | |
| General Feedback: | Both the Pandas DataFrame and Series objects have a robust set of methods available that make them ideal objects for cleaning data. Note that Option 5 occurs “in-place”, which is why you do not need to set the result equal to df[“fare”]. | |
| Hint: | Review Modules 4, 6 and the docs for more on selecting data in Pandas DataFrames: <https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html> | |

**Question 7**

|  |  |  |
| --- | --- | --- |
| Question | Assume the Pandas DataFrame below is assigned to the variable df.  Match the labeled red boxes with the correct values that would display the mean distance traveled for each username in order of largest to smallest distance. If the values are not needed, mark them as “n/a.” |  |
| Image |  |  |
| Option 1 | ‘username’ | **a** |
| Option 2 | ‘distance’ | **b** |
| Option 3 | mean | **c** |
| Option 4 | ascending=False | **d** |
| Option 5 | ‘mean’ | **n/a** |
| Option 6 | ‘driver’ | **n/a** |
| Option 7 | ‘fare’ | **n/a** |
| Option 8 | ascending=True | **n/a** |
| Correct Feedback | Correct! | |
| Incorrect Feedback | Incorrect, please try again. | |
| General Feedback | Grouping and aggregating data is an essential part of exploratory data analysis. | |
| Hint | Grouping and splitting data is covered in module 6-7. | |

**Question 8**

|  |  |  |
| --- | --- | --- |
| Question | Match the code to the distribution plot it produced. | |
| Images |  | |
| Option 1 | mu, sigma = 90, 2  s = np.random.normal(mu, sigma, 10000)  plt.hist(s, bins=20, density=True); | **a** |
| Option 2 | data = np.linspace(0, 10, 100)  fig,ax = plt.subplots(1,1)  ax.plot(data, stats.chi2.pdf(x, df=4)); | **b** |
| Option 3 | data = bernoulli.rvs(size=100,p=0.3)  ax = seaborn.distplot(data, kde=False); | **c** |
| Option 4 | data=binom.rvs(n=10,p=0.7,loc=0,size=1010)  ax=seaborn.distplot(data, kde=False); | **d** |
| Correct Feedback | Correct! | |
| Incorrect Feedback | Incorrect – please try again. See Module 8 for a review on plotting and distributions. | |
| General Feedback | Parametric machine learning models attempt to estimate the parameters of these distributions. | |
| Hint | Reference the module 8-4 to review the shapes of statistical distributions and how to produce them in Python. | |

**Question 9**

|  |  |  |
| --- | --- | --- |
| Question | Consider the DataFrame starships below along with the seaborn heatmap displaying its feature correlations. Which of the following statements are true? Mark all that apply. | |
| Images |  | |
| Option 1 | To add the correlation values to the heatmap, set the parameter annot=True in the seaborn.heatmap function. | **True** |
| Option 2 | The ‘hyperdrive\_rating’ feature is negatively correlated with all other features. | **False** |
| Option 3 | An increase in ‘cost\_in\_credits’ corresponds to an increase in the number of crew members needed to operate the spaceship. | **True** |
| Option 4 | A causal relationship exists between the crew and passenger features. | **False** |
| Option 5 | The seaborn.heatmap() function will compute and display the pairwise column correlations by default, so users do not need to provide the pairwise correlation matrix as input. | **False** |
| Correct Feedback: | Correct! | |
| Incorrect Feedback: | Incorrect – refer to module 11 to exploratory data analysis and plotting. | |
| General Feedback: | This seaborn function is helpful for understanding the relationships between numeric features. The heatmap is especially nice when the data contains a reasonably large number of features. One downside is that it can take a long time to render with large datasets. | |
| Hint: | Refer to module 11 to review this plotting function. The documentation is also helpful: <https://seaborn.pydata.org/generated/seaborn.heatmap.html> | |

**\***Dataset obtained from Kaggle.com: <https://www.kaggle.com/jsphyg/star-wars#starships.csv>

**Question 10**

|  |  |  |
| --- | --- | --- |
| Question | Mark all of the following statements about Maximum Likelihood Estimation (MLE) and Ordinary Least Squares (OLS) as **True** or **False** | |
| Image |  | |
| Option 1 | Both MLE and OLS are probabilistic models for estimating population beta coefficients | **False, OLS isn’t probabilistic** |
| Option 2 | MLE is a method for estimating population parameters, such mean and standard deviation for the normal distribution. The estimated parameters are determined by finding the parameters that would maximize the likelihood of observing the sample data collected. | **True** |
| Option 3 | Reference the two normal distributions A and B plotted above. Distribution B has a higher likelihood than distribution A to produce the following data points:  [ 7, 7, 11, 5, 9] | **False – distribution A is more likely to produce those data points.** |
| Option 4 | OLS almost always performs better than MLE for regression, especially when the assumption of normality is not satisfied. | **False** |
| Correct Feedback: | Correct! | |
| Incorrect Feedback: | Incorrect – refer to module 12-12 to review MLE and OLS. | |
| General Feedback: | MLE and OLS are both legitimate methods for estimating unknown population parameters. The approach should depend on the data and generalized performance. | |
| Hint: | Refer to module 12-12 to review MLE and OLS. | |