Python-worksheet 1

1. C
2. C
3. C
4. A
5. D
6. C
7. A
8. A
9. A
10. C

Statistics- worksheet 1

1. A
2. A
3. B
4. D
5. C
6. B
7. B
8. A
9. C
10. A normal distribution is the proper term for a probability bell curve. In a normal distribution the mean is zero and the standard deviation is 1. Normal distributions are symmetrical, but not all symmetrical distributions are normal.
11. (i) Use deletion methods to eliminate missing data. The deletion methods only work for certain datasets where participants have missing fields. ...

    (ii) Use regression analysis to systematically eliminate data. ...

    (iii) Data scientists can use data imputation techniques.
12. A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.
13. True, imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased.
14. In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).
15. There are three real branches of statistics: data collection, descriptive statistics and inferential statistics.

Machine Learning-worksheet 1

1. A
2. A
3. C
4. B
5. C
6. B
7. D
8. D
9. C
10. A
11. B
12. B, C

13. This is a form of regression, that constrains or shrinks the coefficient estimates towards zero. In other words, this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting. A simple relation for linear regression looks like this. Here 'Y' represents the learned relation and β represents the coefficient estimates for different variables or predictors(X).

    $Y=\beta0+\beta1X1+\beta2X2+\ldots..+\beta pXp$

14. The fitting procedure involves a loss function, known as residual sum of squares or RSS.

    Ridge Regression (L2 Norm) Lasso (L1 Norm) Dropout.
15. Error is the difference between the actual value and Predicted value and the goal is to reduce this difference. ... The blue line is the best fit line predicted by the model i.e the predicted values lie on the blue line