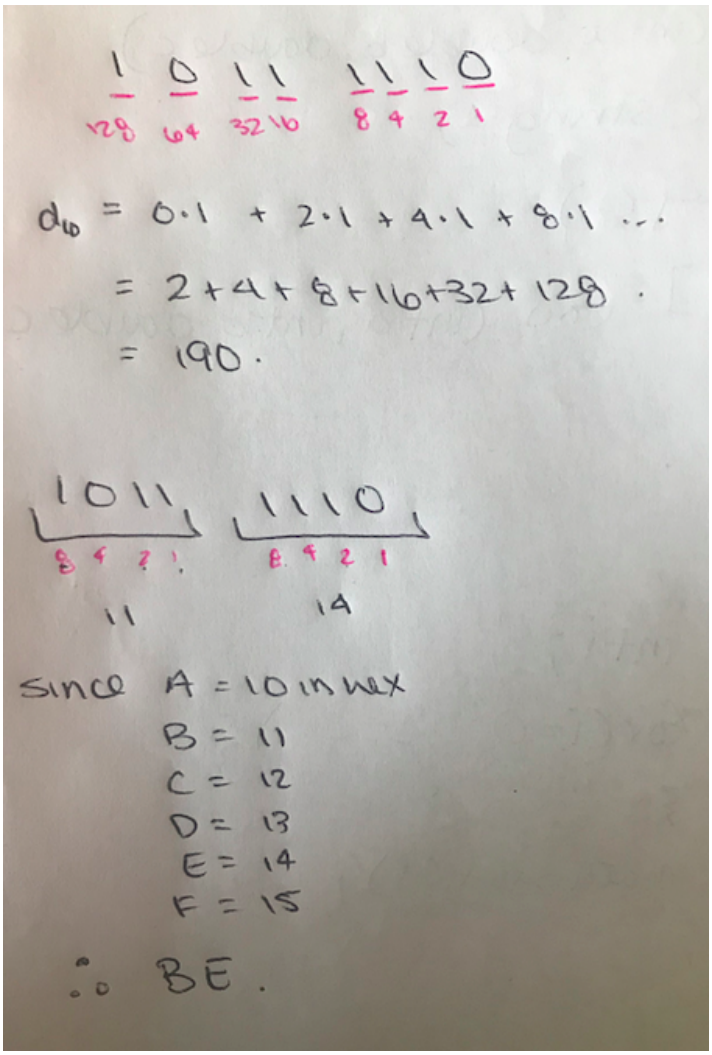
Data 301: Final Review Answers

1. Data analysis is the processing of data to yield useful insights or knowledge. Data Analytics is the science of examining raw data with the purpose of drawing conclusions about that information. A data analyst is a person who uses tools and applications to transform raw data into a form that will be useful. Data analytics is important as society is collecting more and larger data sets all the time: Web All web pages visited and links clicked, searches made, images and posts, Business Items purchased by date, supply chain/customers, industrial sensors, Science Massive data sets (biological/genomic, astronomy, physics, healthcare) Environmental Sensors and monitors (temperature, etc.)

2. Decimal: 190 Hexadecimal: BE

   To get into decimal and hex, I follow the following Procedures:



   Decimal I multiply each, for hex I multiply in groups of 4.

3. D)

A) 10 TB =10,000 GB
B) 100 GB
C) 1,000,000,000,000 bytes = 1000 GB
D) 1 PB = 1,000,000 GB

4. "Big data" is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.
   Data sets grow rapidly, in part because they are increasingly gathered by cheap and numerous information-sensing Internet of things devices such as mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks.[10][11]
   The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s;[12] as of 2012, every day 2.5 exabytes (2.5×1018) of data are generated.[13]
   Based on an IDC report prediction, the global data volume will grow exponentially from 4.4 zettabytes to 44 zettabytes between 2013 and 2020.
   By 2025, IDC predicts there will be 163 zettabytes of data.[14] One question for large enterprises is determining who should own big-data initiatives that affect the entire organization.[15]

5. Computers represent data digitally meaning that data is represented using discrete units called as bits (Binary Digits). The real-world is analog where the information is encoded on a continuous signal (spectrum of values, ie. infinite sounds/colours).
   Example:
   Analog vs. Digital: Thermometer example
   A thermometer contains liquid which expands and contracts in response to temperature changes. The liquid level is analog, and its expansion continuous over the temperature range. This information can be represented using discrete units using digital thermometer, for example.
   Why go Digital?
   1) Computers are digital and many home electronics are interfacing with computers.
   2) Analog signals are more susceptible to noise that degrades the quality of the signal (sound, picture, etc.). The effect of noise also makes it difficult to preserve the quality of analog signals across long distances.
   3) Reading data stored in analog format is susceptible to data loss and noise. Copying analog data leads to declining quality

6. **Metadata** provides the context to understand how to interpret the data to make it useful. Metadata is data that describes other data.
   Examples of metadata include:
   names of files
   column names in a spreadsheet
   table and column names and types in a database

Metadata helps you understand how to interpret and manipulate the data.

7. C) 2
    a. It is possible to have data without metadata.
    b. A character encoded in Unicode uses twice as much space as ASCII.

8. 3
    a. A macro is a recorded set of actions that can be executed repeatedly.
    b. User defined functions can be used in formulas like built in Excel functions.
    c. Macros are only saved in *.xlsm* files.

9. The dollar sign “$” is a symbol that indicates an absolute address. By default, addresses are "relative" in the sense that if they are in a formula that is copied to another cell, they will be changed relative to where they were copied from their origin. However, with absolute addressing, the column or row value will not change when copied from origin.
   See our Midterm 1 review slides for examples.

10. =$B4+G$4

11. AVERAGE(E4:E98)

12. &

13. COUNTIF(A2:A39, “<0”)

14. 3

15. SELECT * FROM emp
    WHERE salary > 20000;

16. SELECT pname
    FROM emp, proj, workson
    WHERE emp.title = 'EE’ AND workson.eno=emp.eno AND workson.pno = proj.pno;

17. error

18. B

19. R is a very powerful statistics language with a large number of libraries for statistical and mathematical calculations.

20. R uses {} and whitespace does not matter. Python is a whitespace language and does not use brackets to denote a suite. Both can have lines terminated with a ; and are case sensitive.

Python will only use = as assignment, while R has a wider (global) scope assignment operator <- . Python indexes from 0, while R indexes from 1.

21. Qualitative data is categorical and refers to groups within data. These are observed, rather than measured. These are often used to denote treatments in experiments.
    Quantitative data is numeric and can be measured.

22. mean: mean()
    median: median()
    variance: var()
    standard deviation: sd()
    range: range()

23. E - all four of them work :)

24. View(dataframe)

25. na.omit() will return a dataframe that has removed any NA values from your original dataframe, reducing the number of rows in the dataframe. is.na() will return a boolean vector that is TRUE for every row in the dataframe that has a NA value.

26. numerically: fivenum(x)
    graphically: boxplot(x)

27. d - both could be zero.

28. You can concatenate a string and a numeric with paste(). This is different from python, where you could convert a numeric to a string with str() and then use +, or simply input both as arguments in print().

29.
```
setwd("crocs")
crocs = read.csv("crocs.csv")
attach(crocs)
t.test(green_crocs, blue_crocs, alternative='greater', paired=FALSE, var.equal=FALSE)
# this is equivalent:
# t.test(green_crocs, blue_crocs, alternative='greater')
```

30. This is more of an understanding question. A paired t test is used when there is reason to believe that the two populations being tested are related. An unpaired t test is used when there is no reason to believe that the two populations are related. In particular, unpaired t tests are used in many comparative experiments. A paired t test needs to have the same number of observations in each vector, while an unpaired does not. So a paired t test is used

to compare the marks on two midterms for one section of a class, while the unpaired t test is used to compare the marks between sections for one midterm.

31. A pvalue is the probability of observing a test statistic as or more extreme than the one observed, assuming the null hypothesis is true.

32.

```
linmod = lm(stress, n_finals)
plot(stress~n_finals,)
abline(linmod$coefficients["(Intercept)"], linmod$coefficients["n_finals"])
# This is a linear model showing the relationship between
# stress levels and the number of finals the students tested have.
# i am so sorry this took me so long, i've been busy.
```

33.

```
new_data = subset(data, variable == 'keep')
```

34. 1
    a. 1:10 creates a vector of ten numbers.

35. 2
    a. Elements in a list may be of different types.
    b. Given matrix m, m[,2] would return all data in column 2.

36. A scalar is a vector of length 1. It is a value on its own, like the number 3.
    A vector is a list of scalars. It is very similar to a list in python, with a key difference being that a vector indexes from 1.
    A matrix is a vector of vectors, or a two dimensional vector. A dataframe will extend on a matrix.

37. You should use *subset* with the optional argument *select*.
    eg: subset(data, condition, select=c(column_1, column_2))

38. Transform is similar to UPDATE in SQL because both of them will change the values of their data structure.
    transform(data, column = column + 1)
    UPDATE data SET column = column + 1

    You can also use transform for log transformations, which are very useful when dealing with non normal data.

39. apply returns a vector or array or list of values obtained by applying a function to margins of an array or matrix.

40. Within the research community, there can be a large pressure to assure that the research yields statistically significant results. For this reason, there can be a large amount of poor statistical practices in the research process. Open Science is a movement aimed at increasing transparency and reducing bias in the scientific community.

41. Any dataset will do, but two examples are: Open Research Datasets by NSERC, Employment Rate Dataset by Statistics Canada.

42. Boolean expression that is either True or False and may contain one or more comparisons.

Syntax	Description
>	Greater than
>=	Greater than or equal
<	Less than
<=	Less than or equal
==	Equal (Note: Not "=" which is used for assignment!)
!=	Not equal

43. Python is a general, high-level programming language designed for code readability and simplicity.
    Python 2 is the legacy of python, while Python 3 is the future of Python. Python 3 is backwards-incompatible meaning Python 3 code won't necessarily run in Python 2.

44. False, True, False, True, False, Error, True

45. print("Data 301 covers a lot of subjects")
    print("Thats okay though because I will rock the exam")
    print("Thanks QSCU for these questions")
    BONUS: print("Data 301 covers a lot of subjects\nThats okay though because I will rock the exam\nThanks QSCU for these questions")
    You can also use '' and '''

46.

```
name = input('what is your name?')
favourite_crocs = input('what is your favourite kind of crocs?')

with open('kat_needs_this_for_science.txt','a') as append_good_crocs:
    append_good_crocs.write(name + ' - ' + favourite_crocs + '\n')


# alt solution
name = input('what is your name?')
favourite_crocs = input('what is your favourite kind of crocs?')

append_good_crocs = open('kat_needs_this_for_science.txt','a')
    append_good_crocs.write(name + ' - ' + favourite_crocs + '\n')
append_good_crocs.close()
```

47.

```
good = True
while good:
    try:
        user_input = input(":)?")
        if user_input == ":(" or user_input == ":/":
            raise Exception(user_input)
    except Exception as e:
        good = False
        print(e)
        print("That's rough buddy")
```

48.

Don't forget to check out our midterm material on our website!
Good luck on your final! You can do it!