2. Data Analytics:

**SQL Query:**

SELECT

Mfg\_length\_data.[Part ID],

Mfg\_length\_data.[Length-Mfg Spec-mm],

Mfg\_length\_data.[Length-Actual-mm],

[Length-Actual-mm]-[Length-Mfg Spec-mm] AS [Signed Error], ([Length-Actual-mm]-[Length-Mfg Spec-mm])/[Length-Mfg Spec-mm]\*100 AS [Percent Signed Error],

Abs([Length-Actual-mm]-[Length-Mfg Spec-mm]) AS [Unsigned Error],

Abs(([Length-Actual-mm]-[Length-Mfg Spec-mm])/[Length-Mfg Spec-mm])\*100 AS [Unsigned Percent Error]

FROM

Mfg\_length\_data

GROUP BY

Mfg\_length\_data.[Part ID],

Mfg\_length\_data.[Length-Mfg Spec-mm],

Mfg\_length\_data.[Length-Actual-mm];

**Mfg\_length\_data worksheet (linked to mfg\_length\_data.accdb)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Part ID** | **Length-Mfg Spec-mm** | **Length-Actual-mm** | **Signed Error** | **Percent Signed Error** | **Unsigned Error** | **Unsigned Percent Error** |
| A04 | 117 | 117 | 0 | 0 | 0 | 0 |
| A22 | 96 | 96 | 0 | 0 | 0 | 0 |
| A27 | 92 | 92 | 0 | 0 | 0 | 0 |
| A18 | 145 | 146 | 1 | 0.689655172 | 1 | 0.689655172 |
| A12 | 112 | 111 | -1 | -0.892857143 | 1 | 0.892857143 |
| A03 | 95 | 96 | 1 | 1.052631579 | 1 | 1.052631579 |
| A20 | 86 | 87 | 1 | 1.162790698 | 1 | 1.162790698 |
| A19 | 83 | 82 | -1 | -1.204819277 | 1 | 1.204819277 |
| A38 | 146 | 144 | -2 | -1.369863014 | 2 | 1.369863014 |
| A09 | 61 | 60 | -1 | -1.639344262 | 1 | 1.639344262 |
| A13 | 61 | 60 | -1 | -1.639344262 | 1 | 1.639344262 |
| A28 | 57 | 56 | -1 | -1.754385965 | 1 | 1.754385965 |
| A30 | 56 | 57 | 1 | 1.785714286 | 1 | 1.785714286 |
| A02 | 98 | 100 | 2 | 2.040816327 | 2 | 2.040816327 |
| A26 | 121 | 124 | 3 | 2.479338843 | 3 | 2.479338843 |
| A21 | 73 | 75 | 2 | 2.739726027 | 2 | 2.739726027 |
| A34 | 59 | 61 | 2 | 3.389830508 | 2 | 3.389830508 |
| A06 | 81 | 78 | -3 | -3.703703704 | 3 | 3.703703704 |
| A10 | 71 | 68 | -3 | -4.225352113 | 3 | 4.225352113 |
| A29 | 144 | 151 | 7 | 4.861111111 | 7 | 4.861111111 |
| A37 | 81 | 77 | -4 | -4.938271605 | 4 | 4.938271605 |
| A36 | 90 | 95 | 5 | 5.555555556 | 5 | 5.555555556 |
| A08 | 142 | 134 | -8 | -5.633802817 | 8 | 5.633802817 |
| A14 | 88 | 83 | -5 | -5.681818182 | 5 | 5.681818182 |
| A16 | 70 | 74 | 4 | 5.714285714 | 4 | 5.714285714 |
| A25 | 90 | 96 | 6 | 6.666666667 | 6 | 6.666666667 |
| A24 | 81 | 87 | 6 | 7.407407407 | 6 | 7.407407407 |
| A11 | 53 | 57 | 4 | 7.547169811 | 4 | 7.547169811 |
| A01 | 126 | 136 | 10 | 7.936507937 | 10 | 7.936507937 |
| A07 | 108 | 117 | 9 | 8.333333333 | 9 | 8.333333333 |
| A05 | 71 | 77 | 6 | 8.450704225 | 6 | 8.450704225 |
| A15 | 69 | 75 | 6 | 8.695652174 | 6 | 8.695652174 |
| A35 | 149 | 136 | -13 | -8.724832215 | 13 | 8.724832215 |
| A23 | 122 | 111 | -11 | -9.016393443 | 11 | 9.016393443 |
| A31 | 54 | 49 | -5 | -9.259259259 | 5 | 9.259259259 |
| A17 | 107 | 117 | 10 | 9.345794393 | 10 | 9.345794393 |
| A39 | 128 | 116 | -12 | -9.375 | 12 | 9.375 |
| A32 | 72 | 79 | 7 | 9.722222222 | 7 | 9.722222222 |
| A33 | 129 | 116 | -13 | -10.07751938 | 13 | 10.07751938 |

a. Calculate the signed errors between the mfg spec and manufactured part lengths. What is the average error?

**Excel**

|  |
| --- |
| Average Error |
| 0.230769231 |

=AVERAGE(Mfg\_length\_data1!D:D)

b. Assuming the mfg spec is the desired length, calculate the unsigned percent error and sort in ascending order. Print the Part ID values associated with the 10 largest percent errors.

|  |
| --- |
| **10 Largest Percent Errors** |
| A33 |
| A32 |
| A39 |
| A17 |
| A31 |
| A23 |
| A35 |
| A15 |
| A05 |
| A07 |

=XLOOKUP(  
LARGE(  
Mfg\_length\_data!$G:$G,  
COUNTA(Mfg\_length\_data!$A$1:A1)  
),  
Mfg\_length\_data!$G:$G,Mfg\_length\_data!$A:$A,"FAIL",0,1)

c. Calculate the standard deviation, σ, of the signed percent error

|  |
| --- |
| **Signed Percent Error Std Dev** |
| 5.745486437 |

=STDEV.P(Mfg\_length\_data!E:E)

d. What percent of the values fall within 1σ, 2σ, and 3σ bounds (looking at percent error)

|  |  |  |
| --- | --- | --- |
| 1 sigma | 2 sigma | 3 sigma |
| 59% | 100% | 100% |

4. Machine Learning (Answer any 3 of the following):

a. Why is reproducibility an important concept in machine learning software development?

b. What methods can be used to identify outliers within a dataset?

c. How do you test whether the data meet the assumptions of linear regression?

d. How are ROC curves and confusion matrices related?

e. Describe your approach to constructing training and test sets. When are validation datasets

appropriate?

f. How does the selection of a regularization method impact model performance?

g. What metrics and visualizations are most helpful when evaluating the performance of a

linear regression model?

5. Binary conversion and memory storage:

a. What is the minimal number of bits needed to represent the decimal 1898 in binary?

0111 0110 1010 = 1898

11 bits can represent the decimal.

b. What is the largest value that can be stored in an unsigned int?

2^BitSize - 1

i.e. 32-bit integer max = 2^32 -1 = 4294967295

c. In what cases would you use a double instead of a float?

When precision is needed or when a wider range of values (larger or smaller than 32-bit floats can provide) is needed, and when memory is not a major concern (double uses twice as much memory as float).

d. Convert 0.1 to binary. What’s a problem with storing certain decimal values in binary?

Certain decimals cannot be represented precisely in binary. As shown below, decimal 0.1 repeats binary decimal places infinitely.

0.1 \* 2

0.2 -> 0.0

0.2 \* 2

0.4 -> 0.00

0.4 \* 2

0.8 -> 0.000

0.8 \* 2

1.6 -> 0.0001

0.6 \* 2

1.2 -> 0.00011

0.2 \* 2

0.4 -> 0.000110

0.4 \* 2

0.8 -> 0.0001100

0.8 \* 2

1.6 -> 0.00011001

0.6 \* 2

1.2 -> 0.000110011

0.2 -> 0.000110011

0.2 \* 2

0.4 -> 0.0001100110

0.4 \* 2

0.8 -> 0.00011001100

…

e. There is a town in Texas where all families have 4 children. Starting with 8 couples, and

assuming monogamous couples which mate for life, what data type and/or how many bits

of memory are necessary to store the number of children at the end of 12 generations.

A screenshot of a table

Description automatically generatedA table of maths

Description automatically generated

Number of 12th gen kids = 65536

Unsigned short int max = 65535

**6. Data Analytics, Experiment Design, and Model Building (Open Short Answer):**

***Prompt***: You're designing a system which processes a wide variety of video data and derives contextual labels and descriptions of the scene. Every input video is accompanied by meta‐data parameters like time of day, camera optics info, geographic location, and others.

Your system has many tuning parameters which define its configuration. Some configurations work better for certain sub‐sets (classes) of the data space, so an ideal system would be smart and adapt based on the input video and accompanying meta‐data. However, neither the classes nor their optimal system configurations are known *apriori*. Due to the many parameters and the range of values they span, the configuration space is too massive to be explored exhaustively. Also, some tuning parameters are more impactful than others, but you don’t necessarily know which ones.

**Question**: How do you approach this problem to design the best performing system? Describe some of the challenges and how you can overcome them. How do you test your system is working?

**Possible Useful Terms**: clustering, optimal, tractability, sensitivity, feasibility, control variable, fix, baseline, experiment, exhaustive, brute‐force search, heuristic, dimensionality, space, scope, PCA, basis, sub‐sampling, classes, combinatorial complexity, correlation, local, global, loss function, residual.

**7. Image Processing:**

a. What are some of the differences in need and application between machine vision imaging cameras and hobbyist photography cameras?

b. You take a picture using a high‐end camera. The raw data is captured and saved along with a compressed (lossy) copy. The compressed copy has several dark areas of the image which seem to lack contrast and look flat, while there are also some pixels in bright areas of the image which exhibit clipping (saturation). What can you do to correct these problems?

c. What are practical challenges one must consider when working with JPG vs PNG files?

d. Looking at the photo of a Triumph TR6 below:

A collage of a car driving on the road

Description automatically generated

Original Mod1 Mod2

Carefully examine the changes in color and lighting which have been introduced from the Original to Mod1. If you were to manually recreate these effects, what specific filters, effects, and methods would you need? What about going from Mod1 to Mod2?