CSSE 386 – Data Mining with Programming
Rose-Hulman Institute of Technology

Worksheet 05

Name (Print):_____ Section:_____

1. **Data manipulation** is the process of _____ or _____ a dataset to explore a research problem.

2. Which research question requires **data manipulation** to answer?

Manufacturer	Model	Drive	EngineType	Cylinders	Liters
Audi	A4	All	Gas	4	2
BMW	328Ci	Rear	Gas	6	3.6
Tesla	Model 3	All	Electric	NaN	NaN
Chevrolet	Malibu	Front	Gas	6	3.6
Ford	Mustang	Rear	Gas	8	5
Rolls-Royce	Ghost	Rear	Gas	12	6.6

   (a) Is there a relationship between a car's gas mileage and engine size?
   (b) Is there a relationship between a car's gas mileage and drive type?
   (c) Is there a relationship between a car's gas mileage and number of cylinders?

3. Which feature(s) could be used to group the car dataset?

   (a) Manufacturer
   (b) Drive
   (c) Liters
   (d) Engine Type
   (e) Model

4. Calculate the group size for:

   - Rear-wheel drive:
   - Front drive:
   - Gas engine type:

5. **Pivot table** is a summarization technique that

   - Groups data based on _____
   - Aggregates data using _____

6. Complete code

```
  # Splits dataframe into subsets
df._____ _______

# Creates pivot tables
df._____________
```

7. Data Summary

```
  # Calculate the mean price of diamonds based on their cut
diamonds.groupby(_____ _______

# Creates the same mean with pivot table
diamonds.pivot_table(_____________

# Creates a query only to select price > 4000
diamonds.query(_____________
```

8. Complete the following in your Colab with titanic dataset from seaborn repository

```
                ## Use groupby()
 # Find out the survival counts (sum) among female and male passengers
 # Change sum to mean
                ## Use pivot_table()
 # Show the average survival rate among male and female passengers
 # Create a bar plot
 # Add columns for classes to see how the survival rate changed across
    classes
 # Create a bar plot
 # Add a query to filter to only look at passengers by themselves (
   check the column Alone) and then look at their survival rate
 # Try another pivot table summary (if time allows)
```

Your observations:

9. **Feature scaling** converts _____ to _____ ranges.

10. Complete table with formula and range

Term	Formula	
Standardization (z-score)		
Normalization (Min-Max)		