ECE M148 Homework 1

Damien Ha

Question 1

In [1]:
```
import statistics
A = [1, 1, 5, 9, 9]
print("The mean is", statistics.mean(A))
print("The median is", statistics.median(A))
```

```
The mean is 5
The median is 5
```

In [2]:
```
B = [1, 1, 5, 9, 9, 11]
print("The mean is", statistics.mean(B))
print("The median is", statistics.median(B))
```

```
The mean is 6
The median is 7.0
```

Overall, A and B seem to have similar centers, with B's mean and median being slightly higher perhaps suggesting B is slightly higher overall. This is consisten with A and B being nearly identical with the exception of B including the value 11 in addition to all the values in A.

Question 2

a.) With random sampling every member of the population has an equal chance of being selected, and it can be very easy to implement which is advantageous. A disadvantage is that, depending on the diversity of the population, the sample may not be representative of the population and there could be bias if the frame of the sample is not representative of the population

b.) Stratified sampling is advantageous in that it ensures relevant subgroups within the population are represented in the data. However, it requires prior knowledge of the population's characteristics and could be time consuming to implement, which could be disadvantageous

c.) Systematic sampling can be more efficient than random sampling and is fairy easy to implement. However, if there's a certain systematic pattern throughout the population, it can result in unrepresentative or biased samples.

d.) Cluster sampling can be effective in gathering good data when the population is very diverse and widepread. Howecer, if the clusters are not well defined then they may not be represetative of the population.

Question 3

1.) One strategy is to completely remove the values from the data, which is quick and easy to implement. However, it may result in a loss of information and bias depending on the randomness of the deleted data.

2.) Another strategy is to replace the missing values with the mean, the mode, or some other value reflective of the data. This strategy should preserve the size of the data and keep the accuracy of the model. However, if the values we replace the nulls with are not reflective of the true value, we may introduce bias.

3.) A third strategy could be to predict the missing values based off the available data using some type of regression, machine learning model, or algorithm, which should help maintain or improve the accuracy and preserve the correlations between variables. However, this takes much more time and computation, and may result in overfitting.

4.) Finally, one could assign "missing" or "unknown" in place of all the null values. Thus, the sample size is maintained and there is no loss of information. However if null values account for a large portion of thr data or if there's some particular pattern in the null values, then this approach would be inappropriate and might not give us accurate information.

Question 4

a.) This is an example of convenience bias. Bob is only asking his wealthy CEO friends who are easily accessible to him as a fellow CEO rather than gathering data that will be representative of the population

b.) This is an example of response bias. Because of the phrasing of Sally's question, students will likely feel pressured to say they're doing well and not want to admit if they're doing poorly or failing, and thus will craft their responses accordingly.

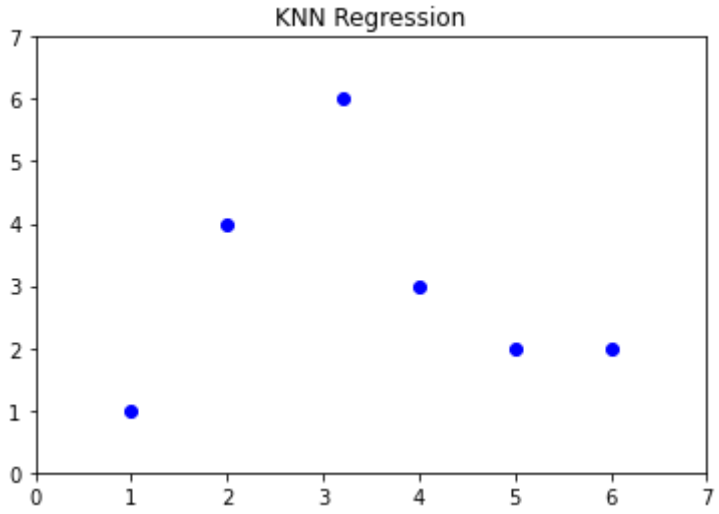
c.) This is an example of voluntary bias. People can choose whether or not to participate in the survey voluntarily which can impact the results. For example, more opinionated people might be more likely to respond, indifferent/ambivalnet people likely won't respond.

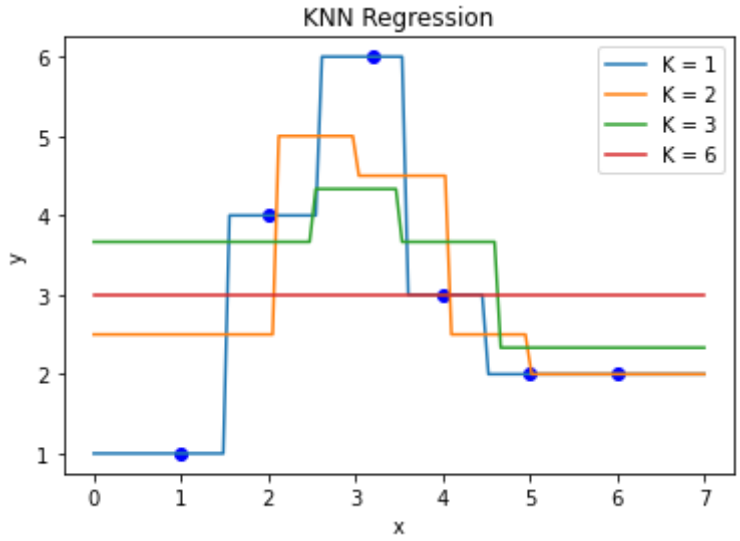
Question 5

In [8]:

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.neighbors import KNeighborsRegressor

X = np.array([[1], [2], [3.2], [4], [5], [6]])
y = np.array([[1], [4], [6], [3], [2], [2]])

plt.xlim(0, 7)
plt.ylim(0, 7)
```

```
plt.scatter(X, y, color='blue')
plt.title('KNN Regression')
plt.show()

X_pred = np.linspace(0, 7, 100).reshape(-1, 1)
k_values = [1, 2, 3, 6]

plt.scatter(X, y, color = 'blue')
for k in k_values:
    knn = KNeighborsRegressor(n_neighbors=k)
    knn.fit(X, y)
    y_pred = knn.predict(X_pred)
    plt.plot(X_pred, y_pred, label='K = {}'.format(k))
plt.legend()
plt.xlabel('x')
plt.ylabel('y')
plt.title('KNN Regression')
plt.show()
```





In this case, the regression line for K=1 follows the training data very closely, whereas larger values like K=6 have a regression line that is much more general and smooth. When K=1, we may overfit the the model whereas with larger K, we will have a more general model that won't be overly influenced by any one point, so K=1 is not necessarily better. However, large K is not always better either. If it is too large, we may not be able to accurately capture important patterns or trends in the data. The optimal K will likely depend on the patterns and complexity of the data we're dealing with.

The KNN performing regression on all $x < 1$ does not seem to be a good idea because this lies outside of the range of our training data, and with KNN we're assuming the new points are similar to the training points. Thus predictions for data outside the range of the training data will be unreliable.