**MKTG5883 Exercise 3: Exploratory Data Analysis**

1. **Exercise 1**
2. The plots are a time-series plot and a sparkline. The sparkline shows exactly the same data, just a more compact form (without the labelling on the axes).
3. Features shown in the data are:
   * A noticeable weekly cycle; probably assignments are due the next day!
   * A sustained, high level of traffic in the first week February - maybe a midterm test.
   * Some days have more than 90 visits, indicating that students visit the site more than once per day, or due to external visitors to the site.
   * **Skewness**: If the histogram is right-skewed, it suggests that most students have relatively low visits, while a few students visit the site very frequently.
   * **Outliers**: The boxplot may reveal outliers (e.g., some students visiting the website an exceptionally high number of times).
4. **Exercise 2**

Histograms are used to show distributions of variables while bar charts are used to compare variables.

Histograms plot quantitative data with ranges of the data grouped into bins or intervals while bar charts plot categorical data.

Bars can be reordered in bar charts but not in histograms.

There are no spaces between the bars of a histogram since there are no gaps between the bins. An exception would occur if there were no values in a given bin but in that case the value is zero rather than a space. On the other hand, there are spaces between the variables of a bar chart.

The bars of bar charts typically have the same width. The widths of the bars in a histogram need not be the same as long as the total area is one hundred percent if percents are used or the total count if counts are used. Therefore, values in bar charts are given by the length of the bar while values in histograms are given by areas.

|  |  |  |
| --- | --- | --- |
| **Feature** | **Bar Chart** | **Histogram** |
| **Data Type** | Categorical | Numerical (Continuous or Interval) |
| **Bar Spacing** | Spaces between bars | No spaces between bars |
| **X-axis** | Categories | Continuous number line or intervals |
| **Bar Represents** | Frequency/Proportion of a category | Frequency/Frequency density within an interval |
| **Order of bars** | can be change, usually sorted | cannot be change, numerical order |

1. **Exercise 3**

* **Data Set:**  
  (1, 1), (1, 2), (2, 1), (2, 1.5), (3, 2), (4, 1.5), (4, 2), (5, 1.5), (4.5, 2), (4, 4), (4.5, 4), (4.5, 5), (4, 5), (5, 5)
* **k = 3** (We want to find 3 clusters)
* **Steps:**
  1. **Initialization:** Randomly choose 3 initial centroids. Let's pick the following points as our initial centroids (this is just one possible starting point; different initializations can lead to slightly different final results):
     + c1 = (1, 1)
     + c2 = (2, 1)
     + c3 = (4, 4)
  2. **Assignment:** Assign each data point to the cluster whose centroid is closest. We use Euclidean distance to measure the distance between a point (x, y) and a centroid (cx, cy):

distance = sqrt((x - cx)^2 + (y - cy)^2)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Point** | **Distance to c1** | **Distance to c2** | **Distance to c3** | **Assigned Cluster** |
| (1, 1) | 0 | 1 | 4.24 | 1 |
| (1, 2) | 1 | 1.41 | 3.61 | 1 |
| (2, 1) | 1 | 0 | 3.61 | 2 |
| (2, 1.5) | 1.12 | 0.5 | 3.20 | 2 |
| (3, 2) | 2.24 | 1.41 | 2.24 | 2 |
| (4, 1.5) | 3.04 | 2.06 | 2.5 | 2 |
| (4, 2) | 3.16 | 2.24 | 2 | 3 |
| (5, 1.5) | 4.03 | 3.04 | 2.69 | 3 |
| (4.5, 2) | 3.64 | 2.69 | 2.06 | 3 |
| (4, 4) | 4.24 | 3.61 | 0 | 3 |
| (4.5, 4) | 4.61 | 4 | 0.5 | 3 |
| (4.5, 5) | 5.32 | 4.74 | 1.12 | 3 |
| (4, 5) | 5 | 4.47 | 1 | 3 |
| (5, 5) | 5.66 | 5.10 | 1.41 | 3 |

* 1. **Update:** Recalculate the centroids of each cluster by taking the mean of all the data points assigned to that cluster.
     + **Cluster 1:** ( (1+1)/2, (1+2)/2 ) = (1, 1.5) New c1
     + **Cluster 2:** ( (2+2+3+4)/4, (1+1.5+2+1.5)/4 ) = (2.75, 1.5) New c2
     + **Cluster 3:** ( (4+5+4.5+4+4.5+4.5+4+5)/8, (2+1.5+2+4+4+5+5+5)/8 ) = (4.44, 3.69) New c3
  2. **Repeat:** Repeat steps 2 (Assignment) and 3 (Update) until the cluster assignments no longer change, or a maximum number of iterations is reached. Let's do one more iteration:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Point** | **Distance to New c1** | **Distance to New c2** | **Distance to New c3** | **Assigned Cluster** |
| (1, 1) | 0.5 | 1.9 | 4.7 | 1 |
| (1, 2) | 0.5 | 2.32 | 4.1 | 1 |
| (2, 1) | 1.12 | 0.86 | 3.7 | 2 |
| (2, 1.5) | 1 | 0.75 | 3.2 | 2 |
| (3, 2) | 2.2 | 0.56 | 2 | 2 |
| (4, 1.5) | 3.0 | 1.25 | 2.2 | 2 |
| (4, 2) | 3.1 | 1.3 | 1.7 | 2 |
| (5, 1.5) | 4.0 | 2.3 | 2.4 | 2 |
| (4.5, 2) | 3.6 | 1.8 | 1.7 | 3 |
| (4, 4) | 4.3 | 3.2 | 0.5 | 3 |
| (4.5, 4) | 4.6 | 3.5 | 0.4 | 3 |
| (4.5, 5) | 5.3 | 4.3 | 1.3 | 3 |
| (4, 5) | 5.0 | 4.0 | 0.8 | 3 |
| (5, 5) | 5.7 | 4.6 | 1.5 | 3 |

Repeat Step 3:

* **Cluster 1:** ( (1+1)/2, (1+2)/2 ) = (1, 1.5) New c1
* **Cluster 2:** ( (2+2+3+4+4+5)/6, (1+1.5+2+1.5+2+1.5)/6) = (3.33, 1.58) New c2
* **Cluster 3:** ( (4.5+4+4.5+4.5+4+5)/6, (2+4+4+5+5+5)/6) = (4.42, 4.17) New c3

| Point | Distance to New c1 | Distance to New c2 | Distance to New c3 | Assigned Cluster |

| (1, 1) | 0.5| 2.4 | 5.3 | 1 |

| (1, 2) | 0.5 | 2.8 | 4.6 | 1 |

| (2, 1) |1.12 | 1.4 | 4.2 | 1 |

| (2, 1.5) | 1 | 1.3 | 3.7 | 1 |

| (3, 2) | 2.2 | 0.6 | 2.5 | 2 |

| (4, 1.5) | 3.0 | 0.7 | 2.7 | 2 |

| (4, 2) | 3.1 | 0.7 | 2.2 | 2 |

| (5, 1.5) | 4.0 | 1.7 | 2.8 | 2 |

| (4.5, 2) | 3.6 | 1.4 | 2.2 | 2 |

| (4, 4) | 4.3 | 2.7 | 0.6 | 3 |

| (4.5, 4) | 4.6 | 2.9| 0.2 | 3 |

| (4.5, 5) | 5.3 | 3.8 | 0.8 | 3 |

| (4, 5) | 5.0 | 3.4 | 0.6 | 3 |

| (5, 5) | 5.7 | 4.1 | 0.9 | 3 |

* **Final Clusters (after a few more iterations):** We continue these iterations. The cluster assignments and centroids will gradually stabilize. The final clusters will look something like this:
  + **Cluster 1:** (1, 1), (1, 2), (2,1), (2, 1.5)
  + **Cluster 2:** (3, 2), (4, 1.5), (4, 2), (5, 1.5), (4.5, 2)
  + **Cluster 3:** (4, 4), (4.5, 4), (4.5, 5), (4, 5), (5, 5)