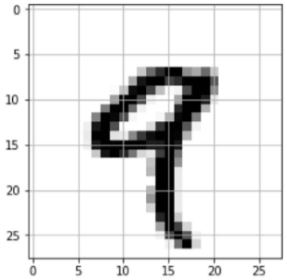
Due on May 1, 2020. Report must show the student's name and USC ID.

1. This exercise is from the book *Hands-on ML with R* by Boehmke. We will work with a data set containing $28 \times 28 = 784$ pixels of $n = 60000$ digits. That is, each image (as shown below) corresponds to a 28-by-28 matrix. Each element in the matrix is a number in [0,255] indicating how dark the pixel is. Each matrix with 784 cells is unscrolled into a row of size 784. The dataset for this question is a dataframe (actually a matrix) of 60000 rows and 784 columns. Think of each row representing a digit as the one shown below.



   a) (20 pts.) We will group the rows into 10 clusters. Then we will compare the clusters with the actual digits. Read Sections 20.1 and 20.5 from

      `https://bradleyboehmke.github.io/HOML/kmeans.html`

      to reproduce and report the output given in that book. Do not use all 60000 rows, instead use a subset of 10000 rows selected at random using `set.seed(1)`.

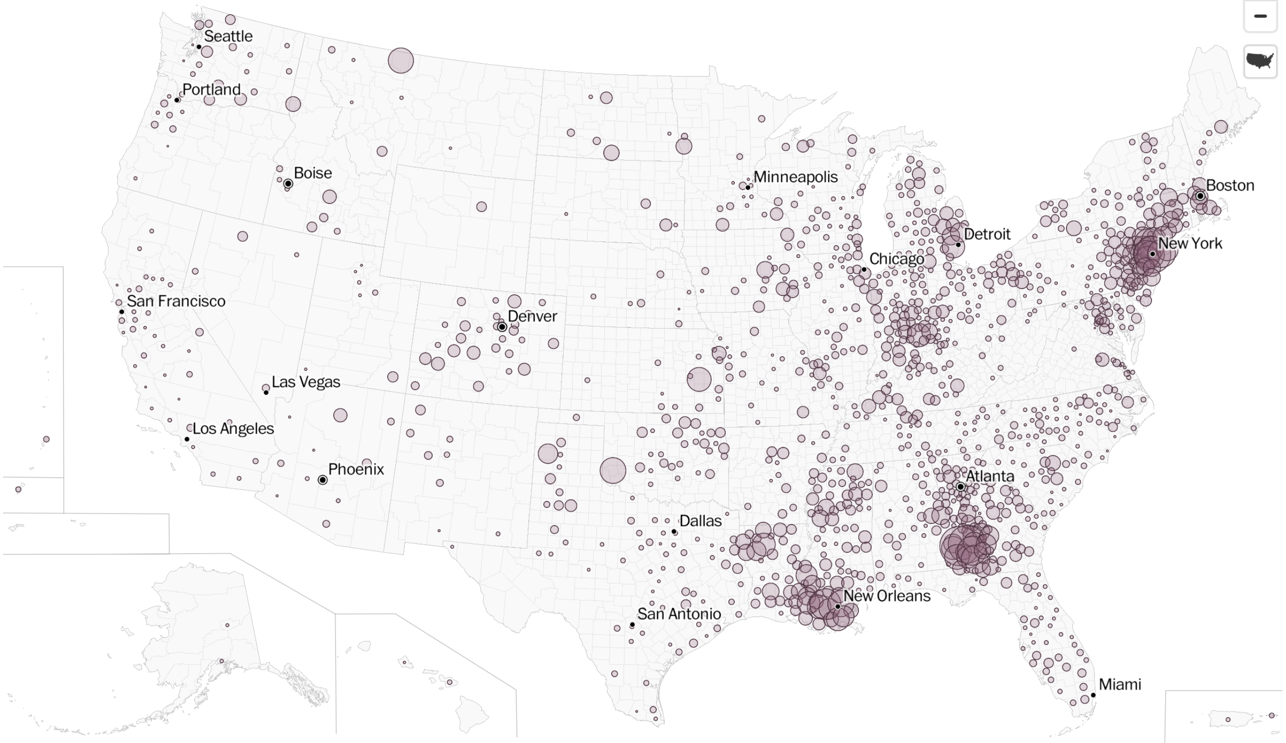
   b) (20 pts.) Use PCA to visualize the data reduced to two dimensions. Create a scatterplot of PC1 vs PC2. Label each point with the actual digit number (different color for each different digit). Use different color for each actual digit. Which digits are well separated? Which are not?

2. The crime dataset from `ggmap` has data from the Houston Police Department over the period of January 2010–August 2010. We are interested in violent crimes that take place downtown. Select `"robbery","aggravated assault","rape", "murder"` categories from column `offense`.

   a) (20 pts.) Create a dot plot showing the location of these offenses in the downtown area (use different dot color for each different offense). Use appropriate legend. Restrict your map to the following coordinates $-95.39681 \leq \text{lon} \leq -95.34188$ and $29.73631 \leq \text{lat} \leq 29.78400$

   b) (20 pts.) Recreate this dot plot using `facet_wrap()` to show the location of each offense in a different facet.

3. (20 pts.) The following map can be found at `https://www.washingtonpost.com` selecting the link `U.S. deaths reported per day` It is useful to visualize the number of coronavirus cases. The size of the circles indicate the number of cases in each location.



The data can be found at https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/
Use `ggmap` to reproduce the map (continental US map only, you may ignore Alaska).

The file `question3.csv` has both the data and the coordinates (lat and lon). It is available on Blackboard. Also, consider using

```
us <- c(left = -125, bottom = 25.75, right = -67, top = 49)
US.map = get_stamenmap(us, zoom = 5, maptype = "toner-lite")
```

(instead of function `get_map`) to get the map.