Assignment 6 WriteUp

Alex Litchfield
12/5/2025
Data Analytics
Assignment 6

Dataset: https://archive.ics.uci.edu/dataset/20/census+income

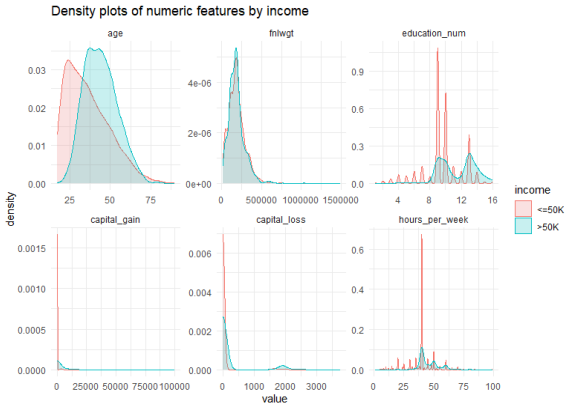
Exploratory Data Analysis (3%):

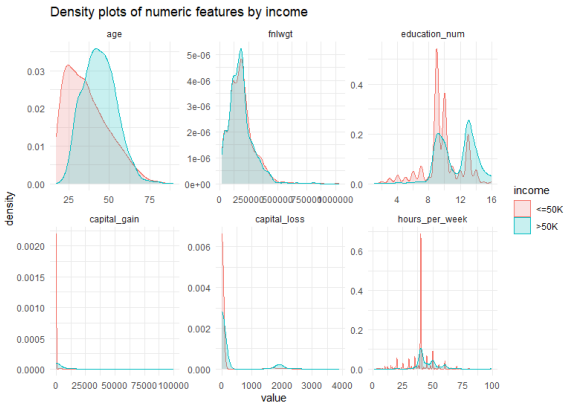
Cleaning and Prepping Data:

The dataset I chose was the Census Income Dataset donated on 4/30/1996. It includes missing values and has many numeric and categorical features. This means that there will need to be different forms of cleaning for each feature. First, the data was already separated into the files *adult.data*, *adult.test*, and *adult.names*. The first 2 files included the training and testing data respectively, and the last one contained the column names. Since each of the first 2 datasets failed to include column names in their files, they needed to be manually added in. Next, I had to tackle the missing values. I found that despite what the dataset claimed, there were no NA values. I later discovered that all missing values were in the form of "?", which meant that I had to convert each "?" value to an NA value and then drop the newly converted NA values. This removed a total of 4,262 from the 32,561 values in the training data and a total of 2,203 from the 16,281 values in the testing data. Finally, I created 2 additional datasets for the training data and testing data that did not include outliers of the sizes 22,209 and 11,114 rows respectively. This filtering was done on the numeric features via the IQR method. I took a random sample of 2,500 rows from the training data with outliers, training data with no outliers, testing data with outliers, and testing data without outliers. This was done after setting a seed to ensure reproducibility.

Data Visualizations and Distributions

Below is the distribution for all numeric features by income (outliers included on the left, no outliers included on the right). Distribution of categorical features can be found on page 7 for both outlier and non-outlier datasets.





Alex Litchfield
12/5/2025
Data Analytics
Assignment 6

Model Development, Validation, and Optimization (7% 6000 Level):

Chosen Models (Classification: Rpart, RF, KNN; Clustering: K-Means):

Since the data included a target feature, categorical features, and numeric features, I decided to perform 3 classification models and 1 clustering model. The first model I used was Rpart. I chose this model because the dataset includes many nonlinear relationships which this model is well equipped for. It also is a solid baseline because it establishes a baseline accuracy percentage for future models. The second model I used was RandomForest. I chose the model because the dataset included many categorical features with different options (for example, 41+ different countries for one feature). This model is good at handling categorical variables with high cardinality, complex nonlinear relationships, and mixed data types, all which fit with the chosen dataset. It also shows how much impact each variable level had on the prediction outcome, which allowed me to remove unnecessary variables from future iterations. The third model I used was KNN. This chose this model because it is a contrasting categorical model than the previous 2, which could provide different results. KNN also performs well when finding local similarity structures in the data, especially when it comes to numeric values. The previous 3 models were classification, which I figured fit well given that the target variable was binary (does the individual make more than $50K annually or do they make less than or equal to $50K annually?). The last model I chose was K-Means Clustering. This model was appropriate because it showed whether certain groups naturally form based on different demographics. It could help illustrate numerical or categorical factors that lead a person to making over $50K a year.

Code Structure (Outliers Included):

The following structure was performed 2 separate times, once with outliers present and once without. First, I ran the 3 classification models and compared their results. Using the RF model, I found that there were 2 features that for the most part did not contribute much to the overall predictions. These were “native country” and “martial status”. Thus, I removed them before moving on to the clustering model. Then I printed the results of the 3 classification models. Upon first try with outliers included, I found that all models exceeded 75% mean accuracy. While this was a great start, I was aiming for higher. Next, I ran all 3 linear models once more with the aforementioned features removed. I found that all 3 models performed nearly identical to the first time, which showed that the features were indeed just noise. Then, I used PCA to prepare the data for K-Means clustering. I created a 2D visualization of the PCA results to find if 2 levels of income were in two tight and distinct clusters. While each point ended up being fairly close to the others in their respective clusters, the clusters had some heavy overlap, indicating that this model may perform poorly. I used the elbow method to find the optimal number of clusters to use, which turned out to be 3. After running the clustering model, I found that the Within-Cluster-Sum-of-Squares (WCSS) values were rather high, indicating that the clusters were very broad. Additionally, the K-Means clustering plot showed 3 clusters which heavily overlapped. It is worth noting that the overlapping clusters, 2 and 3, were the clusters with the highest WCSS values. They also overlap with most of the points in the 2D

visualization of PCA that had the 2 levels of income mixed together. Finally, I ran the 3 classification models a 3rd time using the PCA results to see if those would yield higher model results. This increased the mean accuracy of the KNN model by about 4%, but made the RF and Rpart models perform significantly worse.

Least Contributing Variables - 1st run of RF;

```
native_countryIndia                  3.588106e-01
native_countryMexico                 3.390752e-01
native_countryGreece                 3.212595e-01
native_countryCuba                   3.193708e-01
native_countryIran                   2.264418e-01
native_countryHong                   2.251557e-01
native_countryEcuador                2.101728e-01
education1st-4th                     1.934646e-01
native_countryFrance                 1.739855e-01
native_countryChina                  1.414320e-01
native_countryScotland               1.340307e-01
native_countryJapan                  1.109233e-01
native_countryVietnam                1.070980e-01
native_countryJamaica                9.125484e-02
native_countryEngland                8.661553e-02
native_countryEl-Salvador            8.590645e-02
education5th-6th                     7.725244e-02
native_countryTrinadad&Tobago        7.385564e-02
native_countryNicaragua              6.170167e-02
native_countryYugoslavia             5.482980e-02
native_countryPoland                 5.437668e-02
workclassWithout-pay                 4.460804e-02
native_countryPeru                   2.845184e-02
native_countrySouth                  2.004055e-02
native_countryTaiwan                 1.738276e-02
native_countryLaos                   1.326970e-02
native_countryDominican-Republic     6.824417e-03
native_countryHaiti                  5.627573e-03
educationPreschool                   4.840012e-03
native_countryColumbia               4.598011e-03
occupationPriv-house-serv            3.388008e-03
native_countryGuatemala              3.388008e-03
native_countryIreland                0.000000e+00
native_countryThailand               0.000000e+00
```

Accuracy of all Classification Models (1st Run):

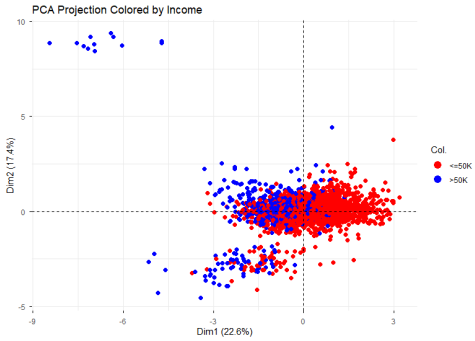
```
Accuracy
       Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
rf    0.834 0.8403194 0.8496994 0.8484038 0.8540000 0.8640000    0
rpart 0.800 0.8080000 0.8100000 0.8160037 0.8263473 0.8356713    0
knn   0.748 0.7500000 0.7655311 0.7595988 0.7660000 0.7684631    0
```

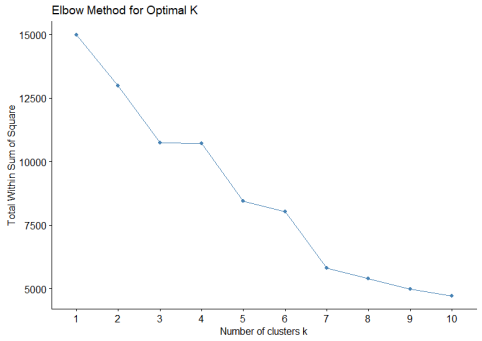
Accuracy of all Classification Models (2nd Run):

```
Accuracy
         Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
rf    0.8323353 0.8336673 0.8440000 0.8448075 0.8483034 0.8657315    0
rpart 0.7935872 0.8123752 0.8180000 0.8143954 0.8203593 0.8276553    0
knn   0.7460000 0.7504990 0.7644711 0.7596088 0.7675351 0.7695391    0
```

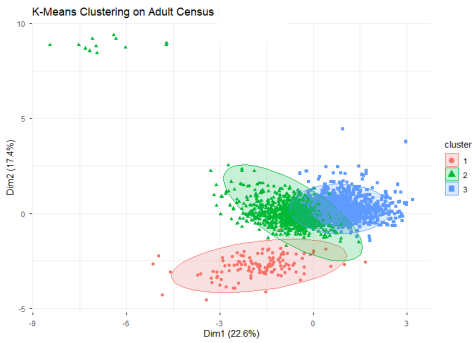
Accuracy of all Classification Models (3rd Run):

```
Accuracy
          Min. 1st Qu.    Median      Mean   3rd Qu.  Max. NA's
rf    0.7780000   0.786 0.7875752 0.7899965 0.7964072 0.802    0
rpart 0.7600000   0.780 0.7900000 0.7856000 0.7920000 0.806    0
knn   0.7784431   0.786 0.7920000 0.7952125 0.8096192 0.810    0
```






```
Within cluster sum of squares by cluster:
[1]  506.4508 6329.8163 3814.1409
 (between_SS / total_SS =  29.0 %)
```



Alex Litchfield
12/5/2025
Data Analytics
Assignment 6

Code Structure (No Outliers Included):

For the second run, I used the dataset with no outliers. First, I ran the 3 classification models and compared their results. Once again, the RF model found that “native country” and "marital status” did not contribute much to the predictions, so they were removed following the first run of the 3 classification models. Then I printed the results of the 3 models, which showed that each model performed within 1% accuracy of the first iteration that included outliers. Following a second run through of each classification model with the unnecessary features removed, the 3 models followed a similar pattern of only moving within 1% accuracy, this time in relation to the first iteration of no outliers. Using PCA, I found that the 2D visualization was nearly identical to a mirrored version of the one generated by the dataset that included outliers. This meant that once again, the clusters were overlapping, though not particularly tight fit. Using the elbow method, I found that the optimal number of clusters was 6, which I used to perform the K-Means clustering model. The WCSS values were high in all 6 clusters except for 1 and 3. Unsurprisingly, these are the 2 clusters which do not show any overlap on the K-Means clustering plot. All other clusters suffered heavy overlap. Finally, I ran the 3 classification models a 3rd time using the PCA results to see if those would yield higher model results. Following the same pattern as before, this increased the mean accuracy of the KNN model by about 4%, but made the RF and Rpart models perform significantly worse.

Least Contributing Variables - 1st run of RF;

raceOther	2.640837e-01
education11th	2.607827e-01
workclassWithout-pay	2.437760e-01
native_countryChina	2.280870e-01
education7th-8th	2.108560e-01
native_countryDominican-Republic	1.958057e-01
education1st-4th	1.228669e-01
native_countryYugoslavia	1.063963e-01
native_countryTaiwan	9.858378e-02
native_countryJapan	8.484158e-02
education9th	7.468535e-02
native_countrySouth	7.434309e-02
native_countryPuerto-Rico	7.331127e-02
occupationPriv-house-serv	7.208279e-02
native_countryNicaragua	5.579817e-02
native_countryIreland	4.084911e-02
native_countryPoland	3.543237e-02
native_countryPeru	3.522620e-02
native_countryEl-Salvador	3.193786e-02
native_countryFrance	1.505369e-02
native_countryEcuador	1.400388e-02
native_countryIran	7.333349e-03
native_countryVietnam	4.956923e-03
native_countryHonduras	4.312457e-03
occupationArmed-Forces	2.560045e-03
native_countryHong	1.706697e-03
native_countryColumbia	1.280023e-03
educationPreschool	0.000000e+00
native_countryGuatemala	0.000000e+00
native_countryOutlying-US(Guam-USVI-etc)	0.000000e+00
native_countryPortugal	0.000000e+00
native_countryThailand	0.000000e+00
native_countryTrinadad&Tobago	0.000000e+00

Accuracy of all Classification Models (1st Run):

Accuracy

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
rf	0.832000	0.8356713	0.836000	0.8387965	0.8443114	0.8460000	0
rpart	0.806000	0.8276553	0.836000	0.8315957	0.8383234	0.8500000	0
knn	0.750499	0.7504990	0.761523	0.7624168	0.7680000	0.7815631	0

Accuracy of all Classification Models (2nd Run):

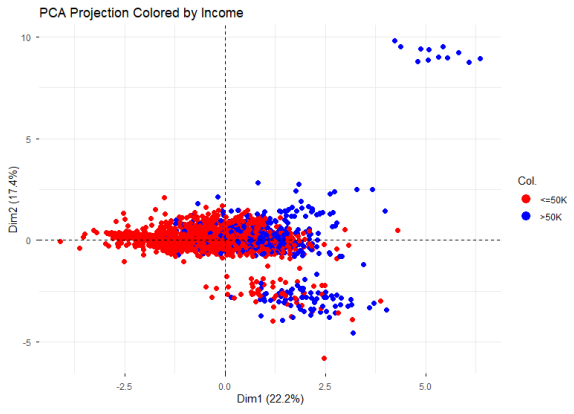
Accuracy

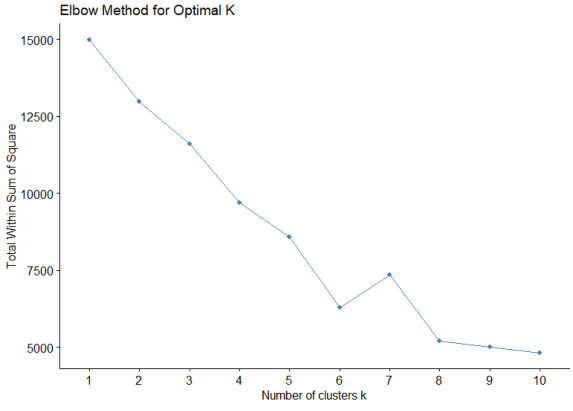
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
rf	0.8200000	0.8336673	0.8440000	0.8403894	0.8440000	0.8602794	0
rpart	0.8076152	0.8180000	0.8276553	0.8259834	0.8343313	0.8423154	0
knn	0.7555110	0.7600000	0.7640000	0.7639932	0.7680000	0.7724551	0

Accuracy of all Classification Models (3rd Run):

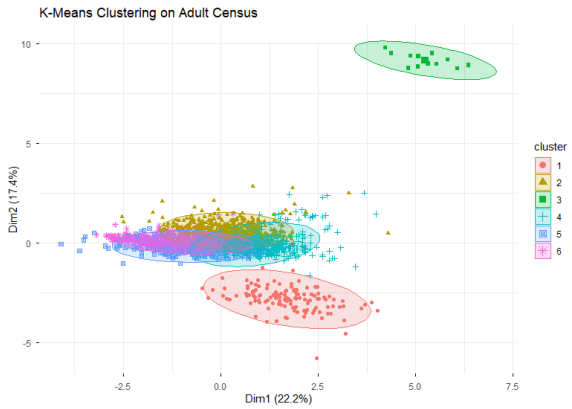
Accuracy

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
rf	0.7780000	0.7924152	0.7955912	0.8020013	0.8160000	0.8280000	0
rpart	0.7675351	0.7784431	0.7860000	0.7895956	0.8000000	0.8160000	0
knn	0.7800000	0.7920000	0.7960000	0.7999917	0.8056112	0.8263473	0





Alex Litchfield
12/5/2025
Data Analytics
Assignment 6




```
within cluster sum of squares by cluster:
[1]  565.34614 1510.40387   33.58581 1747.56928  946.83385 1486.06368
 (between_SS / total_SS =  58.1 %)
```

Model Development, Validation, and Optimization (5% 6000 Level):

Classification Models / Data supplied	Mean Accuracy (With Outliers)	Mean Accuracy (Without Outliers)
Random Forest (All data)	84.8%	83.9%
Rpart (All data)	81.6%	83.1%
KNN (All data)	76.0%	76.2%
Random Forest (Features Removed)	84.5%	84.0%
Rpart (Features Removed)	81.4%	82.6%
KNN (Features Removed)	76.0%	76.4%
Random Forest (PCA)	79.0%	80.2%
Rpart (PCA)	78.6%	79.0%
KNN (PCA)	79.5%	80.0%

Conclusion:

The results give me both confidence in the classification models and the Census Income dataset. Across all models and data variations, a constant pattern emerged, which was that income is not uniformly distributed across variables. Instead, it seems to be associated with a smaller number of predictors, which is why I was able to filter some out with little to no change in the results. The output provided by Random Forest reinforces this theory by showing the exact breakdown for how dominant each contributing variable was. This shows that the data does have relationships between variables.

Alex Litchfield
12/5/2025
Data Analytics
Assignment 6

Random Forest performed the best among all the models, likely due to its ability to handle heavy processing and capture nonlinear relationships. It was able to effectively handle the mix of categorical and numerical features. Rpart, while having a slightly worse accuracy overall, followed this same pattern. KNN performed the worst among the 3 classification models, but this was expected due to the large amount of variation in each categorical feature (some having 40+ options). The feature reduction that was performed had little to no impact on the models, but reinforced the output from Random Forest that the “native country” and “marital status” features were not necessary for the predictions and likely just noise.

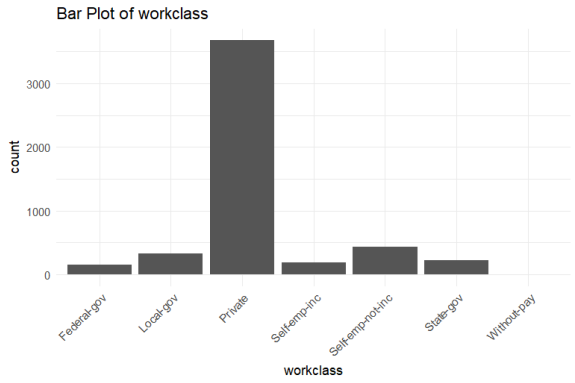
The clustering and PCA, while not showing clear and distinct variation among clusters, illustrated that there are some combinations of variables which point to a predictable income. The different K values (k=3 and k=6) showed that the natural groupings were not just random, as the optimal K value only changed when outliers were removed. Without these extreme values, it gave the model a better chance at predicting the non-outlier income levels. Since the clustering model did not perform overly well, it is possible that there was too much noise from some of the included features. Alternatively, there could be additional factors that determine income which were not present in this dataset.

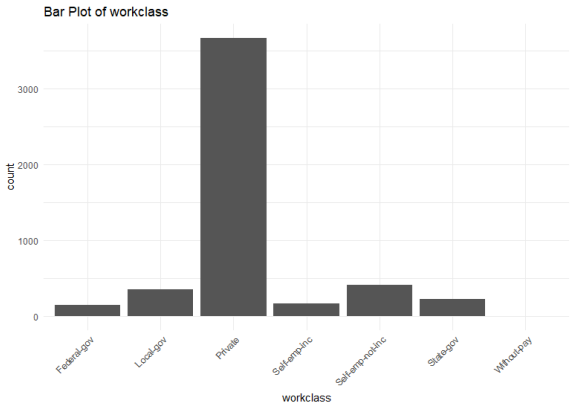
Applying the classification models to PCA produced slightly lower accuracy for the tree based models, reinforcing the idea that PCA may remove important structure needed for prediction. While PCA was able to increase the overall performance of the KNN model, it is clear that this was not because the PCA performance itself was beneficial for the model. Instead, it created a streamlined accuracy percentage across the board that happened to exceed what the base KNN model could accomplish.

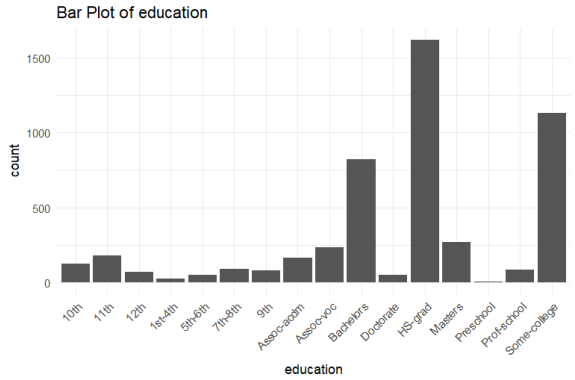
Overall, Random Forest is the best model to forward with if further analysis were to be needed. It provides the best variation in terms of its ability to handle different data types and the highest accuracy percentage. It also was key in showcasing unnecessary features that were included in the original dataset. These models show that predicting income from these features is feasible and could be used in real world decision making.

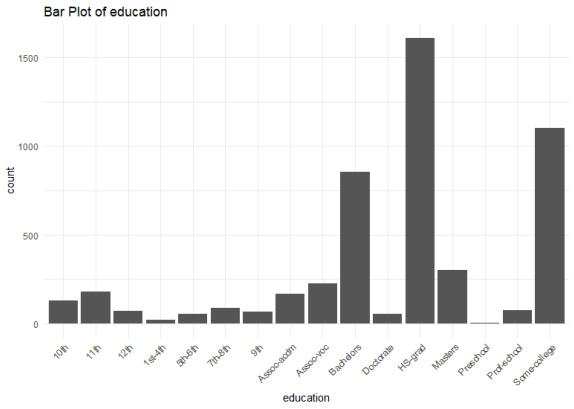
Categorical Plots EDA

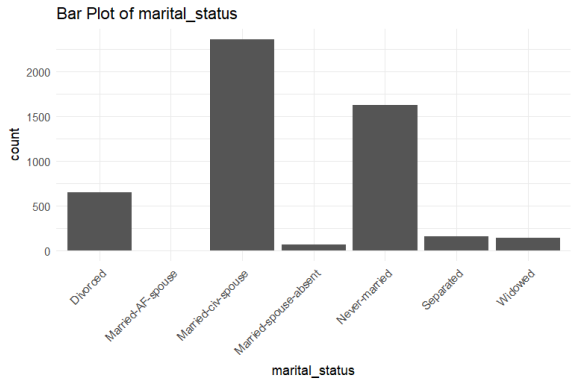
Figures on left (Outliers Included); **Figures on right (Outliers NOT Included)**

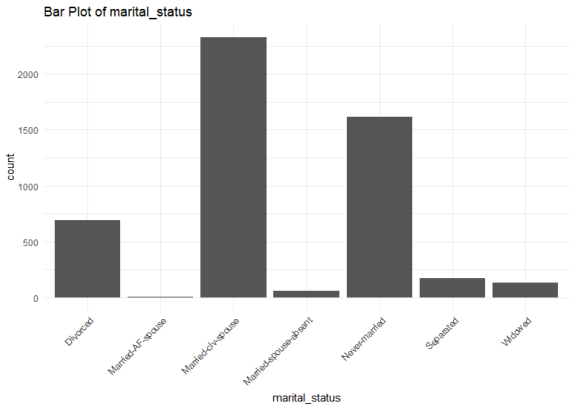


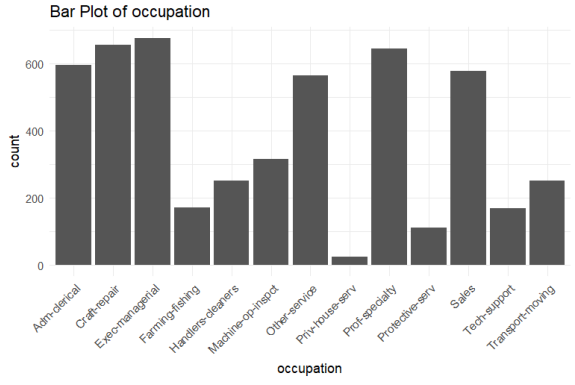


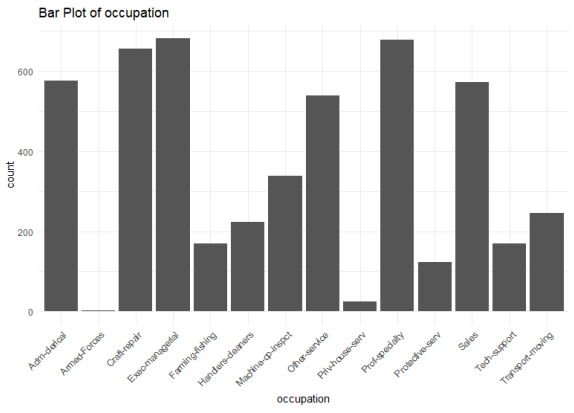


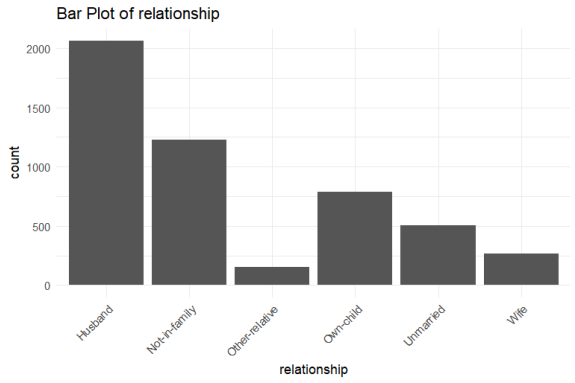


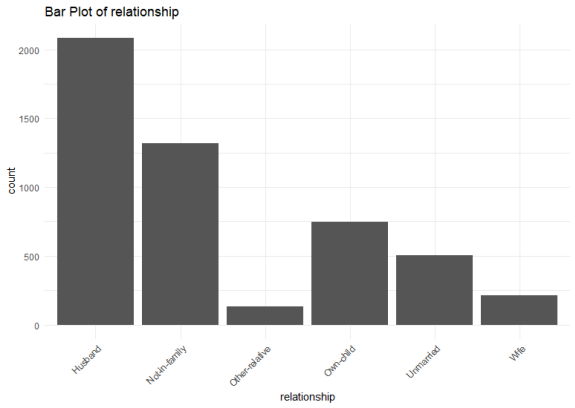


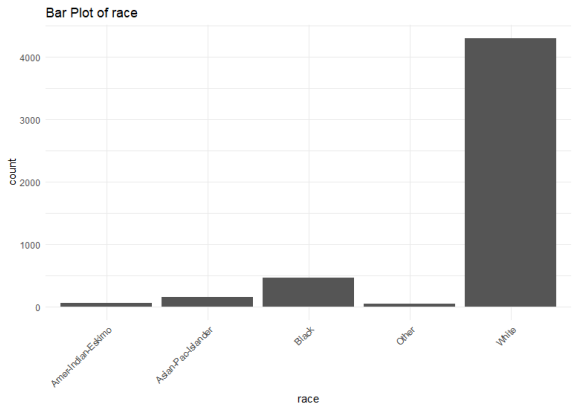


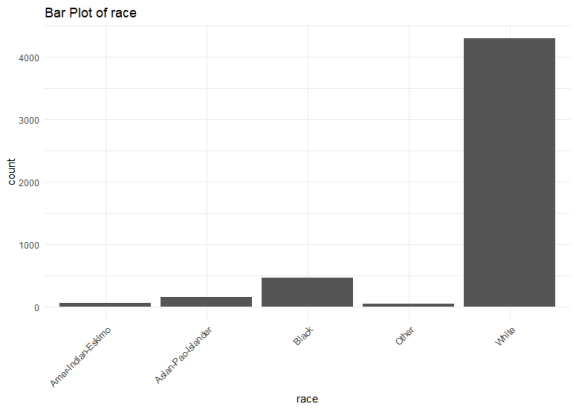


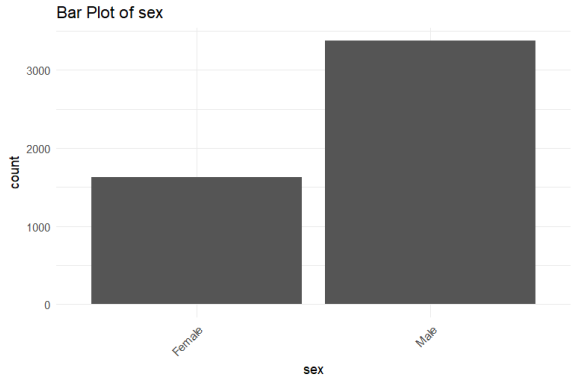


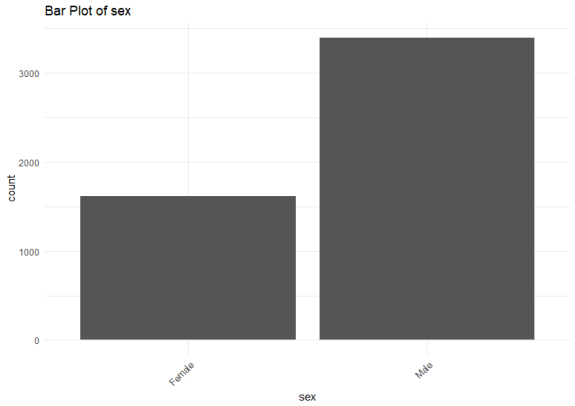

Bar Plot of relationship
count
relationship
Husband
Not-in-family
Other-relative
Own-child
Unmarried
Wife

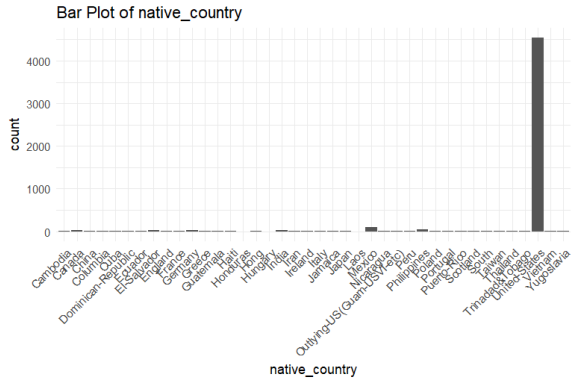

Bar Plot of relationship
count
relationship
Husband
Not-in-family
Other-relative
Own-child
Unmarried
Wife

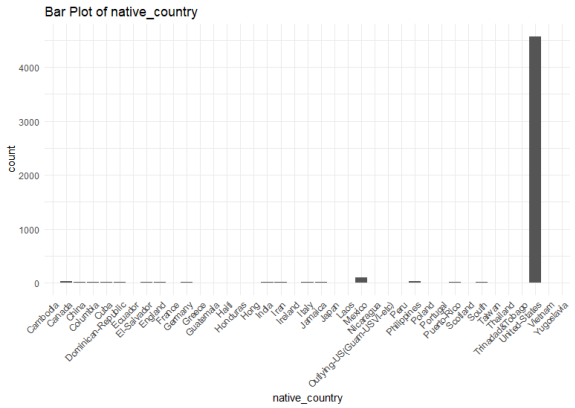

Bar Plot of race
count
race
Amer-Indian-Eskimo
Asian-Pac-Islander
Black
Other
White

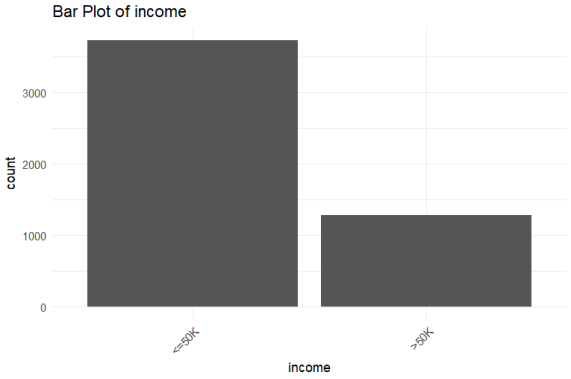

Bar Plot of race
count
race
Amer-Indian-Eskimo
Asian-Pac-Islander
Black
Other
White

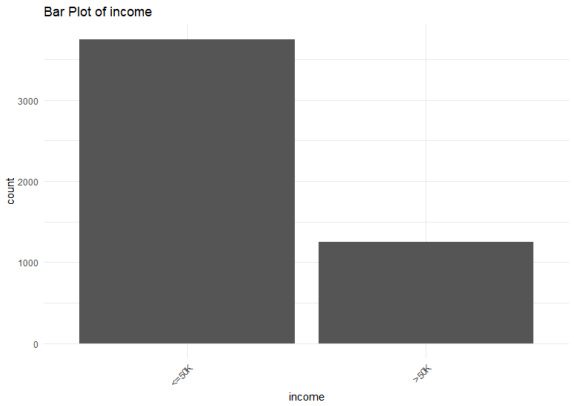

Bar Plot of sex
count
sex
Female
Male


Bar Plot of sex
count
sex
Female
Male


Bar Plot of native_country
count
native_country


Bar Plot of native_country
count
native_country


Bar Plot of income
count
3000
2000
1000
0
<=50K
>50K
income


Bar Plot of income
count
3000
2000
1000
0
<=50K
>50K
income