**Data Story Part:**

**Project overview:**

Breast cancer is the second most prevalent cancer diagnosed among US women, accounting for nearly one in three cancers. It is also the second leading cause of cancer death among women after lung cancer. The two most commonly used screening methods for breast cancer, physical examination of the breasts and mammography, offer an approximate likelihood that a lump is cancer. When these examinations are inconclusive, a procedure known as fine needle aspiration is performed to help establish the diagnosis. This procedure involves removing a sample of fluid in the lump and the sample is investigated under microscope. The physician assesses the health of the sample based on various features of the cells (e.g. area, radius, texture, and so forth) visible in microscopic images. The objective of the procedure is to determine if the suspicious mass is benign (not life threatening) or malignant (cancerous) and correct diagnosis at this stage could lead to implementing effective treatment to either decelerate or stop the cancer. However, this procedure lacks the repeatability required for a firm statement on the disease stage. Advancements in computer analysis of microscope images and development of powerful machine learning methods could significantly improve the accuracy of breast cancer diagnosis. Therefore, the objective of the current project is to develop predictive models that could distinguish benign and malignant cases based on features computed from a digitized image of a fine needle aspirate of a breast mass.

**Breast cancer Wisconsin data set:**

The dataset is published by Kaggle and taken from the University of California Irvine (UCI) machine learning repository. Features are computed from a digitized image of a fine needle aspirate of a breast mass. They describe characteristics of the cell nuclei present in the image.

Attribute Information:

1) ID number 2) Diagnosis (M = malignant, B = benign) 3-32)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter) b) texture (standard deviation of gray-scale values) c) perimeter d) area e) smoothness (local variation in radius lengths) f) compactness (perimeter^2 / area - 1.0) g) concavity (severity of concave portions of the contour) h) concave points (number of concave portions of the contour) i) symmetry j) fractal dimension ("coastline approximation" - 1)

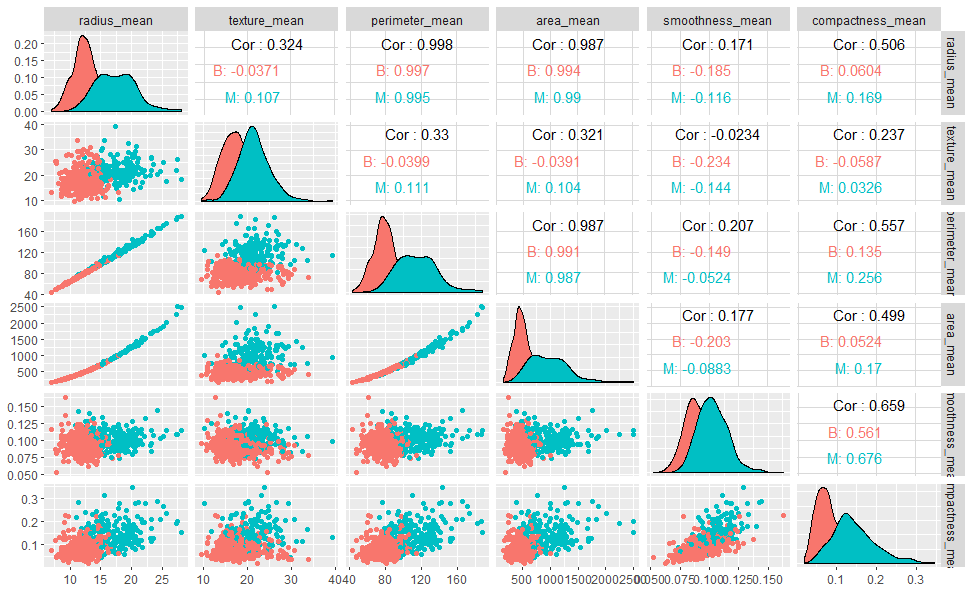
The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. All feature values are recoded with four significant digits.

**Data wrangling**

The last column of the data set was removed as it included an unknown variable with NA values. Additionally, the column for patient ID was removed as it contained no useful information for our analysis. Other than that, the data set was tidy and consisted of no missing values.

**Data exploration**

Visualization below shows the pairwise scatter plots of some of the selected features (mean radius, texture, perimeter, area, compactness, and smoothness). Number of plotted features was limited in order to keep the plot readable. The red and green points correspond to benign and malignant cases, respectively. The data suggests the existence of highly correlated variables in the data set which should be paid attention in development of regression models to avoid collinearity. The plot illustrates the fact that features in malignant cases have higher values than benign cases and this suggests the predictive potential of these features for classification of benign and malignant cases.



**Data analysis**

This data set consists of features that showed promise to distinguish between benign and malignant cases. These features combined with advanced machine learning models could be a more robust and reliable method of diagnosis than conventional diagnostic methods.

This should be noted that because of highly correlated variables in this data set, an essential step in development of regression model would be identifying the best subset of variables that could offer a simple model with minimal collinearity and acceptable accuracy in prediction.

This data lacks the information about the progression of breast cancer in benign and malignant cases. Therefore, some questions such as “when/what percentage of benign cases develop breast cancer in future?” or “when is the best time for intervention in malignant cases?” could not be answered with this data set.

**Future step**

The next step in the project will be exploring the machine learning algorithms that could be used to develop models for diagnosis. Metrics like accuracy, sensitivity, specificity, and misclassification error could be utilized to evaluate the performance of each of the models.