* [Overview](#gjdgxs)
* [Tips from Johnny](#30j0zll)

# CSE 450 - Machine Learning & Data Mining

## Module 02 — Workplace Makeover, Major Case Study Assignment

## Overview

After a few more meetings with the executive team, the head of the data science division has assigned your team to address the following issues asked by the stakeholders:

##### 

Cecil, the VP of Human Resources asks:

There are two main things we need to know. First, is there a pay gap based on gender or race for similar positions and experience levels?

Second, how are we doing on employee retention? We've had some uh—less than stellar methods of trying to measure that in the past.

##### 

William, the VP of Finance asks:

We spent a lot of money hiring the excellent Employee Evaluators Inc. to conduct an employee engagement survey. I know it's a great company, because it's run by my brother, Harold.

Cecil seems to think that her own homegrown survey is better. I admit hers was cheaper..., but sometimes you get what you pay for. For some reason, Devon thinks you can tell us which survey is a better indicator of employee longevity.

I guess we'll see.

##### 

Cecil, the VP of Human Resources adds:

Yes, thank you William.

So what we're looking for specifically is, if we put employees into groups based on their months of service, do we see any correlation between how long they've been here and the results of the latest satisfaction survey I did (or the engagement study done by William's brother)?

Can we identify what the members of each group have in common?

##### 

Devon, the CEO asks:

These are all great questions, thanks team!

One last thing. The most important question to the board right now is diversity. Are we doing a good job on diversity overall? If not, which of our recruiting sources is having the biggest benefit? Where should we be focusing future recruitment efforts?

If you could send us [your write up on this by Saturday night](http://docs.google.com/summary.docx), that would be great.

##### 

Johnny, the data science intern, whispers to you after the meeting:

Just to clarify something Cecil said, in the case of terminated employees, their months of service would be the time between their start date and their termination date. For active employees, it would be the time between their start date and today.

Also, I put together a list of tips and ideas that might help us out:

## Tips from Johnny

Data Dictionary

Our database analyst put together this [data dictionary](http://docs.google.com/hr-dictionary.txt) to help explain the values and sources of different columns in the [hr dataset](https://raw.githubusercontent.com/byui-cse/cse450-course/master/data/hr.csv), so be sure to review that.

Salary Scale

Remember that the PayRate column is using two different scales. Hourly employees (production technicians) have their hourly rates listed. Salaried employees (everyone else) have their annual salaries listed.

If you're doing any kind of comparison of pay rates, you'll need to convert annual salaries to hourly rates, remembering that salaried employees work 40 hours a week, 52 weeks a year.

Feature Scaling

If you're going to be comparing different numeric features, be sure they are using the same scale. You may find it useful to use min-max scaling to handle this problem. You could do this calculation manually, or use [Sci-Kit Learn's MinMaxScaler](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html)

Binning

Just as you did with the Titanic dataset when you reduced the number of titles, you may find it useful to "bin" categorical features into discrete groups in order to address some of the questions above. There are multiple ways to do this, but previously we used the [map() function](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.map.html).

Dealing with Dates

In order to do calculations on dates, you'll first need to convert the feature from a string to a DateTime object. The [to\_datetime() function](https://pandas.pydata.org/pandas-docs/stable/getting_started/intro_tutorials/09_timeseries.html) can help with this.

Once the feature is a DateTime, you can [calculate the number of months between two dates](https://medium.com/@bramtunggala/a-simple-way-to-finding-the-difference-between-two-dates-in-pandas-179d2714b6c).

K-Means Clustering

Remember, K-Means doesn't work on categorical data. According to what I got from the discussion, you'll be clustering based on months of service and the results of the old employee engagmeent survey and Cecil's most recent employee satisfaction survey.

I'm sure you know this, but you'll need to tackle this part in two steps. First, find the optimal number of clusters using the elbow method. Then, cluster the data into that many clusters and try to figure out what the members of each cluster have in common with one another.

Visualizations

Aside from the visualizations you'll make for the other questions, you will probably want to visualize your clusters in two different ways:

1. A scatter plot visualization where each point is colored according to its cluster label.
2. One or more scatter plots where each point is colored according to some feature you think is important to the clustering.

Alternatively if the important feature is categorical, you could do a single chart where the color is based on cluster label, and the size of the point is based on the feature you think is important to the clustering.

##### 

Johnny, the Data Science Intern, drops by your house around midnight:

Okay, just one last thing, if you need any more help at all, I put together [this collection of Google Colab notebooks](http://docs.google.com/hints.html) that might be useful.

Also, your cat looked hungry, so I fed him. Hope you don't mind.

1. [CEO photo by Oz Seyrek on Unsplash](https://unsplash.com/photos/-Ir03_pgpMU)  [↩](#1fob9te)
2. [VP of HR photo by Christina @ wocintechchat.com](https://unsplash.com/photos/SJvDxw0azqw) [↩](#3znysh7)
3. [VP of Finance photo by steffen Wienberg on Unsplash](https://unsplash.com/photos/ml-pxK0Ovmw) [↩](#2et92p0)
4. [Data Science Intern photo by Fábio Lucas on Unsplash](https://unsplash.com/photos/iczrMDNuvzkml-pxK0Ovmw) [↩](#tyjcwt)