# **EXPLORATORY DATA ANALYSIS OF THE HUMAN RESOURCE DATASET**

**1.Outline**

**2.Introduction**

**3.Report**

**4.Proposed Model to Cater for Summary Statistics**

**5.Abstract**

**6.Data Understanding**

**7.Data Analysis and Preparation**

**8.Descriptive Statistics**

**9.Alternative Model Proposed Model to Cater for Gender Diversity**

# **Report**

Your goal is to be critical of the report that was given to you by another team member. You need to consider ways in which this analysis could be improved and provide your own interpretation of the situation before this report is shared with company leadership. The questions listed in each heading are suggestions. You do not need to answer each one, and your report can explore other questions not listed here.

**Business Understanding**

### **Abstract**

In the last decades, having the best machines was enough to be competitive or to dominate an industrial sector. Nowadays, the company that has more engaged and productive employees will have a better chance of winning market competition. For this reason, companies can not lose important employees and when that begins to happen you need to understand why, to prevent this from happening. The Human Resources Analytics dataset, is used to explain the first steps in the data analysis path. In this first part is presented how to get familiarize itself with the data set by performing the descriptive analysis. Techniques such as exploratory data analysis (EDA) allow us to present the data in a more meaningful way, applying general statistical methods and exploratory graphics, that allow a simpler interpretation before engage a machine learning algorithm.

**Data Understanding.**

-Is the data appropriate? What don’t we know from the data that would be helpful when understanding the results? What data should they have included that was missing?

○ How effective are the visualizations at building the narrative of the report? How could they be improved? What visualizations are missing that could help?

## **Exploratory Data Analysis (EDA)**[**¶**](http://localhost:8888/notebooks/human-resources-analytics-a-descriptive-analysis.ipynb#2.-Exploratory-Data-Analysis-(EDA))

Exploratory data analysis employs a variety of techniques (mostly statistical graphics) before making inferences from data. It is essential to examine all variables in the dataset to:

* Catch mistakes
* Generate hypotheses
* See patterns in the data
* Extract important variables
* Detect outliers and anomalies
* Gain deep familiarity with the dataset
* Refine selection of features that will be used to build the machine learning models.

Special attention to not skip the EDA process, because can generate inaccurate models or accurate models on the wrong data. This dataset contains 14999 objects and 10 attributes described below:

Variables | Descriptions

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

satisfaction\_level | Satisfaction Level

last\_evaluation | Last evaluation

number\_project | Number of projects

average\_montly\_hours | Average monthly hours

time\_spend\_company | Time spent at the company

Work\_accident | Whether they have had a work accident

left | Whether the employee has left

promotion\_last\_5years | Whether had a promotion in the last 5 years

sales | Departments (column sales)

salary | Salary

## **3. Preprocessing the dataset**

Before starting the process, its important to answer if it's clear what kind of problem we are dealing with, because in many cases isn't so simple to identify it. A good understanding of the problem will help to choose the right data mining and machine learning techniques to make the right predictions.

Thus, the first step, is preprocessing the data to look for missing, incomplete or noise values, because, in real word, the raw data can be collect from many sources like sensors, websites, public data and many others.

-start the step of preprocessing the dataset is necessary to import some useful Python libraries.

* Numpy: Is a fundamental package to use linear algebra and random number capabilities. See: [www.numpy.org/](http://www.numpy.org/)
* Pandas: Is a package to work with relational data as tables. See: pandas.pydata.org/

**Data Preparation**

-Did the report appropriately explore all the different ways in which the data may

be corrupted? What were additional cleaning steps they could have considered? Should they have reshaped the data in any way? Do you trust the data? What would make you trust or distrust the data?

**Modeling**

-Were the analytics choices here appropriate? Did they apply them correctly? You don’t need to know the specifics of the code, but more about the general approach (e.g. was a decision tree a good choice, or is there another analysis that would have been better used?)

○ How were the models evaluated? How do you know that they fit the data appropriately? What approaches did they use to avoid overfitting? Do we know if these models will work on unknown data in the future? What metrics could they have used to assess the quality of the model?

○ What were the important variables in the models and how do you know they are important? Do we know how these variable impact the outcome? How could they have measured that impact?

**Evaluation**

**Deployment**

# Human Resources Analytics: A Descriptive Analysis

### Introduction

In the last decades, having the best machines was enough to be competitive or to dominate an industrial sector. Nowadays, the company that has more engaged and productive employees will have a better chance of winning market competition. For this reason, companies can not lose important employees and when that begins to happen you need to understand why, to prevent this from happening. The Human Resources Analytics dataset, is used to explain the first steps in the data analysis path. In this first part is presented how to get familiarize itself with the data set by performing the descriptive analysis. Techniques such as exploratory data analysis (EDA) allow us to present the data in a more meaningful way, applying general statistical methods and exploratory graphics, that allow a simpler interpretation before engage a machine learning algorithm.

## 1. The Human Resources Dataset

The Human Resources Analytics is a simulated dataset from [Kaggle](https://www.kaggle.com/ludobenistant/hr-analytics) and the focus is to understand why the best and most experienced employees is leaving the company. By the exploration of this dataset its possible to extract good insights of a problems that the Human Resource department deals daily. In many industries retain their best employees its a question of long term strategy, and can impact the companies growth or put in financial risk, mainly if the employees leave to work at the competitor.

## 2. Exploratory Data Analysis (EDA)

Exploratory data analysis employs a variety of techniques (mostly statistical graphics) before making inferences from data. It is essential to examine all variables in the dataset to:

* Catch mistakes
* Generate hypotheses
* See patterns in the data
* Extract important variables
* Detect outliers and anomalies
* Gain deep familiarity with the dataset
* Refine selection of features that will be used to build the machine learning models.

Special attention to not skip the EDA process, because can generate inaccurate models or accurate models on the wrong data. This dataset contains 14999 objects and 10 attributes described below:

Variables | Descriptions

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

satisfaction\_level | Satisfaction Level

last\_evaluation | Last evaluation

number\_project | Number of projects

average\_montly\_hours | Average monthly hours

time\_spend\_company | Time spent at the company

Work\_accident | Whether they have had a work accident

left | Whether the employee has left

promotion\_last\_5years | Whether had a promotion in the last 5 years

sales | Departments (column sales)

salary | Salary

## 3. Preprossesing the dataset

Before starting the process, its important to answer if it's clear what kind of problem we are dealing with, because in many cases isn't so simple to identify it. A good understanding of the problem will help to choose the right data mining and machine learning techniques to make the right predictions. Thus, the first step, is preprocessing the data to look for missing, incomplete or noise values, because, in real word, the raw data can be collect from many sources like sensors, websites, public data and many others.

To start the step of pre-processing the dataset is necessary to import some useful Python libraries.

* Numpy: Is a fundamental package to use linear algebra and random number capabilities. See: www.numpy.org/
* Pandas: Is a package to work with relational data as tables. See: pandas.pydata.org/

In [1]:

**import** **numpy** **as** **np**

**import** **pandas** **as** **pd**

#### Load the data

To load the dataset we use a Pandas method called **read\_csv** that read CSV(comma-separated) files and covert into DataFrame.

In [2]:

data = pd.read\_csv('../input/HR\_comma\_sep.csv')

Other useful method is **info** that shows a summary of the dataset, like number of observations, columns, variable type and the total memory usage. The dataset have 14999 observations, 10 columns and with no null values. The data types of the variables are divided in 2 float, 6 integer and 2 object.

In [3]:

data.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 14999 entries, 0 to 14998

Data columns (total 10 columns):

satisfaction\_level 14999 non-null float64

last\_evaluation 14999 non-null float64

number\_project 14999 non-null int64

average\_montly\_hours 14999 non-null int64

time\_spend\_company 14999 non-null int64

Work\_accident 14999 non-null int64

left 14999 non-null int64

promotion\_last\_5years 14999 non-null int64

sales 14999 non-null object

salary 14999 non-null object

dtypes: float64(2), int64(6), object(2)

memory usage: 1.1+ MB

Let's see the first 5 lines of the dataset. The **head** method list first N rows from the DataFrame and the method **tail**, returns the last N rows.

In [4]:

data.head(5)

Out[4]:

|  | **satisfaction\_level** | **last\_evaluation** | **number\_project** | **average\_montly\_hours** | **time\_spend\_company** | **Work\_accident** | **left** | **promotion\_last\_5years** | **sales** | **salary** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 0.38 | 0.53 | 2 | 157 | 3 | 0 | 1 | 0 | sales | low |
| **1** | 0.80 | 0.86 | 5 | 262 | 6 | 0 | 1 | 0 | sales | medium |
| **2** | 0.11 | 0.88 | 7 | 272 | 4 | 0 | 1 | 0 | sales | medium |
| **3** | 0.72 | 0.87 | 5 | 223 | 5 | 0 | 1 | 0 | sales | low |
| **4** | 0.37 | 0.52 | 2 | 159 | 3 | 0 | 1 | 0 | sales | low |

In [5]:

data.tail(5)

Out[5]:

|  | **satisfaction\_level** | **last\_evaluation** | **number\_project** | **average\_montly\_hours** | **time\_spend\_company** | **Work\_accident** | **left** | **promotion\_last\_5years** | **sales** | **salary** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **14994** | 0.40 | 0.57 | 2 | 151 | 3 | 0 | 1 | 0 | support | low |
| **14995** | 0.37 | 0.48 | 2 | 160 | 3 | 0 | 1 | 0 | support | low |
| **14996** | 0.37 | 0.53 | 2 | 143 | 3 | 0 | 1 | 0 | support | low |
| **14997** | 0.11 | 0.96 | 6 | 280 | 4 | 0 | 1 | 0 | support | low |
| **14998** | 0.37 | 0.52 | 2 | 158 | 3 | 0 | 1 | 0 | support | low |

**sample** is a easy way to get a few data quickly.

In [6]:

data.sample(10)

Out[6]:

|  | **satisfaction\_level** | **last\_evaluation** | **number\_project** | **average\_montly\_hours** | **time\_spend\_company** | **Work\_accident** | **left** | **promotion\_last\_5years** | **sales** | **salary** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **6047** | 0.98 | 0.55 | 3 | 260 | 3 | 0 | 0 | 0 | sales | medium |
| **9153** | 0.41 | 0.45 | 3 | 236 | 2 | 0 | 0 | 0 | RandD | medium |
| **2870** | 0.75 | 0.79 | 4 | 145 | 3 | 0 | 0 | 0 | technical | medium |
| **3142** | 0.81 | 0.80 | 4 | 229 | 2 | 0 | 0 | 0 | sales | medium |
| **14786** | 0.39 | 0.49 | 2 | 129 | 3 | 0 | 1 | 0 | technical | medium |
| **11860** | 0.78 | 0.90 | 5 | 158 | 3 | 0 | 0 | 0 | sales | low |
| **13773** | 0.76 | 0.79 | 3 | 247 | 3 | 0 | 0 | 0 | sales | low |
| **3454** | 0.89 | 0.86 | 3 | 178 | 4 | 0 | 0 | 0 | sales | medium |
| **10941** | 0.81 | 0.92 | 5 | 258 | 3 | 0 | 0 | 1 | sales | medium |
| **12274** | 0.11 | 0.90 | 6 | 254 | 4 | 0 | 1 | 0 | technical | low |

#### Variables transformations

To plot some statistical graphics and for better understanding, we make some transformations in the variables:

* sales: Rename to department
* salary: Convert the type of the variable from categorical to numerical.

In [7]:

*# RENAME column sale to department*

data.rename(columns={'sales': 'department'}, inplace = **True**)

*# Convert salary variable type to numeric*

data['salary'] = data['salary'].map({'low':1, 'medium':2, 'high':3})

## 4. Descripitve Analysis

The descripitve Analysis is used to simplify and summarize the mainly characteristics of the dataset. In other words, show what kind of information the dataset has. The Pandas method **describe** generates a descriptive statistics that summarize the central tendency, dispersion and shape of the dataset. By using this method in Human Resource dataset important insights is possible to see:

* That approximately 24% os the employees left the company.
* The satisfaction level is around 62% and performance is around 72%.
* Employees work in average on 4 projects with 200 hours worked per month.

In [8]:

data.describe()

Out[8]:

|  | **satisfaction\_level** | **last\_evaluation** | **number\_project** | **average\_montly\_hours** | **time\_spend\_company** | **Work\_accident** | **left** | **promotion\_last\_5years** | **salary** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Count** | 14999.000000 | 14999.000000 | 14999.000000 | 14999.000000 | 14999.000000 | 14999.000000 | 14999.000000 | 14999.000000 | 14999.000000 |
| **Mean** | 0.612834 | 0.716102 | 3.803054 | 201.050337 | 3.498233 | 0.144610 | 0.238083 | 0.021268 | 1.594706 |
| **Std** | 0.248631 | 0.171169 | 1.232592 | 49.943099 | 1.460136 | 0.351719 | 0.425924 | 0.144281 | 0.637183 |
| **Min** | 0.090000 | 0.360000 | 2.000000 | 96.000000 | 2.000000 | 0.000000 | 0.000000 | 0.000000 | 1.000000 |
| **25%** | 0.440000 | 0.560000 | 3.000000 | 156.000000 | 3.000000 | 0.000000 | 0.000000 | 0.000000 | 1.000000 |
| **50%** | 0.640000 | 0.720000 | 4.000000 | 200.000000 | 3.000000 | 0.000000 | 0.000000 | 0.000000 | 2.000000 |
| **75%** | 0.820000 | 0.870000 | 5.000000 | 245.000000 | 4.000000 | 0.000000 | 0.000000 | 0.000000 | 2.000000 |
| **max** | 1.000000 | 1.000000 | 7.000000 | 310.000000 | 10.000000 | 1.000000 | 1.000000 | 1.000000 | 3.000000 |

### 4.1 How many employees works in each department?

Depending on how many employees work in each department, you can learn more about the type of company segment.

In [9]:

print(data['department'].value\_counts())

sales 4140

technical 2720

support 2229

IT 1227

product\_mng 902

marketing 858

RandD 787

accounting 767

hr 739

management 630

Name: department, dtype: int64

### 4.2 How many employees per salary range?

The employees salary is divided in Low (1), Medium (2) and High (3), distributed as follows:

In [10]:

print(data['salary'].value\_counts())

1 7316

2 6446

3 1237

Name: salary, dtype: int64

### 4.3 How many employees per salary range and department?

In [11]:

table = data.pivot\_table(values="satisfaction\_level", index="department", columns="salary",aggfunc=np.count\_nonzero)

table

Out[11]:

| **salary** | **1** | **2** | **3** |
| --- | --- | --- | --- |
| **department** |  |  |  |
| **IT** | 609.0 | 535.0 | 83.0 |
| **RandD** | 364.0 | 372.0 | 51.0 |
| **accounting** | 358.0 | 335.0 | 74.0 |
| **hr** | 335.0 | 359.0 | 45.0 |
| **management** | 180.0 | 225.0 | 225.0 |
| **marketing** | 402.0 | 376.0 | 80.0 |
| **product\_mng** | 451.0 | 383.0 | 68.0 |
| **sales** | 2099.0 | 1772.0 | 269.0 |
| **support** | 1146.0 | 942.0 | 141.0 |
| **technical** | 1372.0 | 1147.0 | 201.0 |

### 4.4 How plot graphics?

In descriptive analysis is very useful to use graphics to represent the data. For that, is necessary to import the libraries:

* Matplotlib: is a plotting library, useful to plot statistical graphics. See: www.matplotlib.org
* Seaborn: is a library based on matplotlib that can draw attractive statistical graphics. See: seaborn.pydata.org/index.html

In [12]:

%**matplotlib** inline

**import** **matplotlib.pyplot** **as** **plt**

**import** **seaborn** **as** **sns**

sns.set()

#### Boxplot

A boxplot is a way of summarizing a set of data measured on an interval scale. It is often used in exploratory data analysis. It is a type of graph which is used to show the shape of the distribution, its central value, and variability. The picture produced consists of the most extreme values in the data set (maximum and minimum values), the lower and upper quartiles, and the median. [Definition taken from Valerie J. Easton and John H. McColl's Statistics Glossary v1.1]

**Figure 1.** Boxplot example

Boxplot is a good statistical graphic to analyze the dataset and identify outliers values. An outlier is as observation that lies an abnormal distance from other values, in this case the analyst have to decide what is considered abnormal.

The boxplots below, give the information about the data distributions:

* Satisfaction level and Last evaluation has a skewed left (negative) distributions.
* Number of projects has a skewed right(positive)distribution.
* Average monthly hours has a symmetric distribution.

Analyse de distribution of the variables is important due the fact that many statistical tests assume normal distribution.

In [13]:

f, axes = plt.subplots(2,2, figsize=(10,10), sharex=**True**)

plt.subplots\_adjust(wspace=0.5)*# adjust the space between the plots*

sns.despine(left=**True**)

*# plot a boxplot of satisfaction\_level to see if there is outliers*

sns.boxplot( x= 'satisfaction\_level', data=data, orient='v',ax=axes[0,0])

*# plot a boxplot of last\_evaluation to see if there is outliers*

sns.boxplot( x= 'last\_evaluation', data=data, orient='v',ax=axes[0,1])

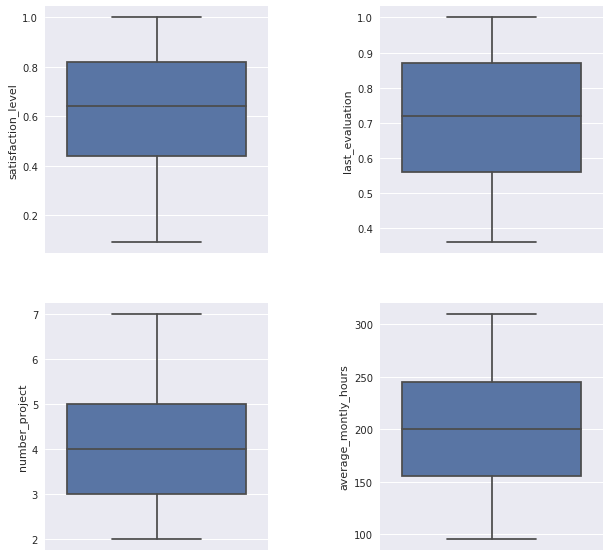
*# plot a boxplot of number\_project to see if there is outliers*

sns.boxplot( x= 'number\_project', data=data, orient='v',ax=axes[1,0])

*# plot a boxplot of average\_montly\_hours to see if there is outliers*

sns.boxplot( x= 'average\_montly\_hours', data=data, orient='v',ax=axes[1,1]);

*#Put a ; at the end of the last line to suppress the printing of output*



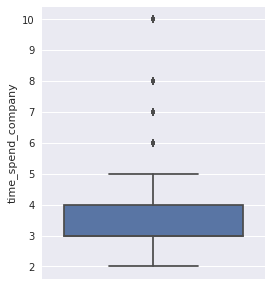
In the boxplots below it is possible to see that only time\_spend\_company has outliers. Let's explain what kind of information is possible to conclude:

* The employees with more time in the company have 10 years, so is possible to say that is a relatively young company.
* Most of the employees have between 3 or 4 years in the company.

In [14]:

plt.figure(figsize=(4,5))

sns.boxplot( x= 'time\_spend\_company', data=data, orient='v');



### 4.5 Correlation Analysis

The correlation is a very useful statistical analysis that describes the degree of relationship between two variables. Let´s see the table below and the heat map to see what relationship are in the data.

In the heat map is possible to see:

* Negative correlation of (-0.39) between satisfaction\_level and the employees that left the company.
* The highest positive correlation is between number of projects and average monthly hours (0.42).
* Last\_evaluation is high correlated to number\_project(0.35)and average\_monthly\_hours(0.34).
* Work\_accident have a low negative correlation(-0.15)and salary (-0.16) with employees that left.

In [15]:

corr = data.corr()

corr

Out[15]:

|  | **satisfaction\_level** | **last\_evaluation** | **number\_project** | **average\_montly\_hours** | **time\_spend\_company** | **Work\_accident** | **left** | **promotion\_last\_5years** | **salary** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **satisfaction\_level** | 1.000000 | 0.105021 | -0.142970 | -0.020048 | -0.100866 | 0.058697 | -0.388375 | 0.025605 | 0.050022 |
| **last\_evaluation** | 0.105021 | 1.000000 | 0.349333 | 0.339742 | 0.131591 | -0.007104 | 0.006567 | -0.008684 | -0.013002 |
| **number\_project** | -0.142970 | 0.349333 | 1.000000 | 0.417211 | 0.196786 | -0.004741 | 0.023787 | -0.006064 | -0.001803 |
| **average\_montly\_hours** | -0.020048 | 0.339742 | 0.417211 | 1.000000 | 0.127755 | -0.010143 | 0.071287 | -0.003544 | -0.002242 |
| **time\_spend\_company** | -0.100866 | 0.131591 | 0.196786 | 0.127755 | 1.000000 | 0.002120 | 0.144822 | 0.067433 | 0.048715 |
| **Work\_accident** | 0.058697 | -0.007104 | -0.004741 | -0.010143 | 0.002120 | 1.000000 | -0.154622 | 0.039245 | 0.009247 |
| **left** | -0.388375 | 0.006567 | 0.023787 | 0.071287 | 0.144822 | -0.154622 | 1.000000 | -0.061788 | -0.157898 |
| **promotion\_last\_5years** | 0.025605 | -0.008684 | -0.006064 | -0.003544 | 0.067433 | 0.039245 | -0.061788 | 1.000000 | 0.098119 |
| **salary** | 0.050022 | -0.013002 | -0.001803 | -0.002242 | 0.048715 | 0.009247 | -0.157898 | 0.098119 | 1.000000 |

In [16]:

sns.set(style='white')

mask = np.zeros\_like(corr, dtype=np.bool)

mask[np.triu\_indices\_from(mask)] = **True**

*# Inserir a figura*

f, ax = plt.subplots(figsize=(13,8))

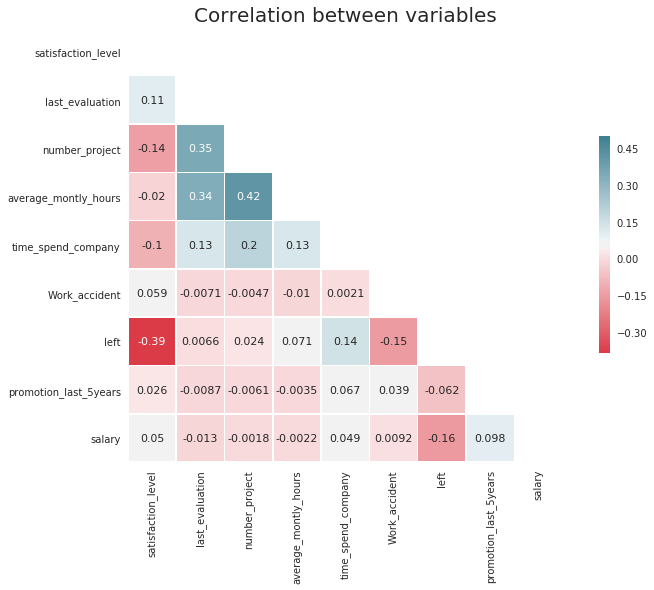
cmap = sns.diverging\_palette(10,220, as\_cmap=**True**)

*#Desenhar o heatmap com a máscara*

ax = sns.heatmap(corr, mask=mask, cmap=cmap, vmax= .5, annot=**True**, annot\_kws= {'size':11}, square=**True**, xticklabels=**True**, yticklabels=**True**, linewidths=.5,

cbar\_kws={'shrink': .5}, ax=ax)

ax.set\_title('Correlation between variables', fontsize=20);



## 5. Hypothesis

Now let's extract some more information and testing some hypothesis

### 5.1 How many employees left the company?

In [17]:

print(data['left'].value\_counts()[1],"employees left the company")

3571 employees left the company

In [18]:

*# The plot show the amount o employees that stayed and left the company.*

plt.figure(figsize=(4,5))

ax = sns.countplot(data.left)

total = float(len(data))

**for** p **in** ax.patches:

height = p.get\_height()

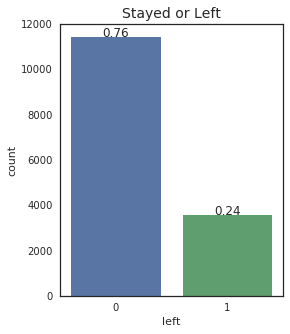
ax.text(p.get\_x()+p.get\_width()/2.,

height + 3,

'**{:1.2f}**'.format(height/total),

ha="center")

plt.title('Stayed or Left', fontsize=14);



### First Hypothesis

The first hypothesis is that salary is the reason why the employees left the company. Let's see if is this correct.

In [19]:

j = sns.factorplot(x='salary', y='left', kind='bar', data=data)

plt.title('Employees that left by salary level', fontsize=14)

j.set\_xticklabels(['High', 'Medium', 'Low']);



In the graphic Salaries by department is possible to see the distribuition of the salaries by department.

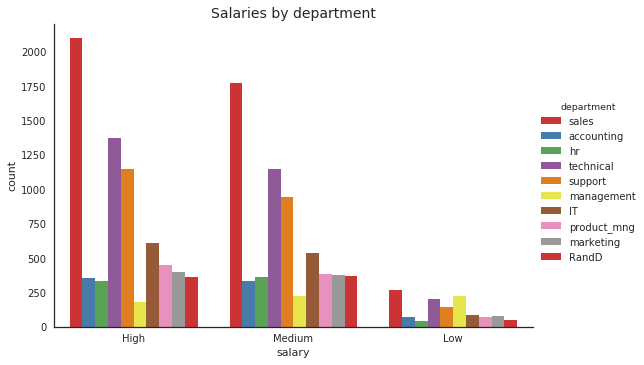
* Most of the employees of the sales department have low or medium salaries, this may be due that in some companies the sales commission is paid separately.
* Technical department is in the second place where most of the employees receives low and medium salaries.

In [20]:

h = sns.factorplot(x = 'salary', hue='department', kind ='count', size = 5,aspect=1.5, data=data, palette='Set1' )

plt.title("Salaries by department", fontsize=14)

h.set\_xticklabels(['High', 'Medium', 'Low']);



In the graphic(Salary Comparison):

* The management department has the biggest difference between the salary of the employees who stayed and those that left.
* It's not possible to see a huge difference in other departments.

The first hypothesis looks very weak to be the main reason why the employees left the company.

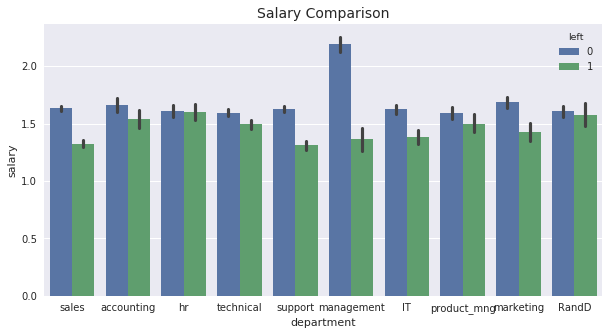
In [21]:

sns.set()

plt.figure(figsize=(10,5))

sns.barplot(x='department', y='salary', hue='left', data=data)

plt.title('Salary Comparison', fontsize=14);



### Second Hypothesis

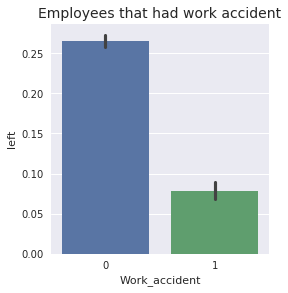
It is a dangerous job?

The second hypothesis is: employees leave the company because work is not safe.

In [22]:

sns.factorplot(x='Work\_accident', y='left', kind='bar', data=data)

plt.title('Employees that had work accident', fontsize=14);



About 14% of the employees had a work accident, although of the high number only of accidents only 169 employees data left the company had work a accident. Then this hypothesis is discarded.

In [23]:

print(data.Work\_accident.sum())

print(data.Work\_accident.mean())

print((data[data['left']==1]['Work\_accident']).sum())

2169

0.1446096406427095

169

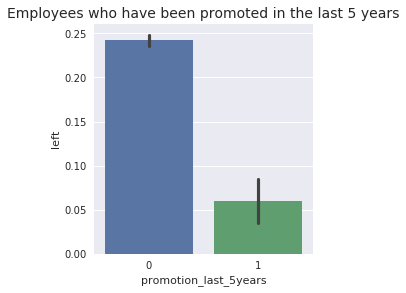
### Third Hypothesis

Is this company a good place to grow professionally?

In [24]:

sns.factorplot(x='promotion\_last\_5years', y='left', kind='bar', data=data)

plt.title('Employees who have been promoted in the last 5 years', fontsize=14);



In the last five years only 319 employees had promotion, this is equivalent to 2% of all employees. This may be a problem because if it is difficult to get promoted many employees become unmotivated and start looking for a new job.

In [25]:

print(data.promotion\_last\_5years.sum())

print(data.promotion\_last\_5years.mean())

319

0.021268084538969265

**Years in the company**

In the graphic 'Years in the company' we can identify an important characteristic.

* Employees with 7 or more years didn't left, maybe because with the passing of the years they are more confortable and not so interested in look for a new challenge in other company.
* The problem starts when the employees have more than 3 years and get worst when they achieve 5 years.
* It is too early to say that the difficult to get promoted is the main reason for the leaving of the employees, but more research is needed.

In [26]:

plt.figure(figsize =(7,5))

bins = np.linspace(1.0, 11,10)

plt.hist(data[data['left']==1]['time\_spend\_company'], bins=bins, alpha=1, label='Employees Left')

plt.hist(data[data['left']==0]['time\_spend\_company'], bins=bins, alpha = 0.5, label = 'Employee Stayed')

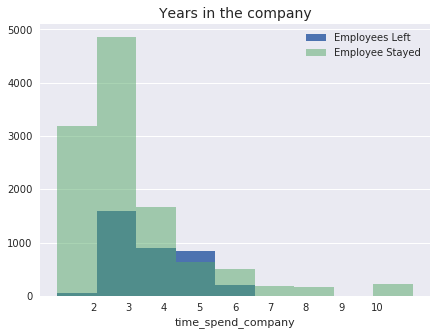
plt.grid(axis='x')

plt.xticks(np.arange(2,11))

plt.xlabel('time\_spend\_company')

plt.title('Years in the company', fontsize=14)

plt.legend(loc='best');



### **Performance Analysis**

There are 2 distinct groups of employees. A group with poor performance and other with high performance employees. It's natural that employees that don't work well leave the company, but the main problem is that the high-performance employees is leaving too and it's necessary to understand why.

In [27]:

plt.figure(figsize =(7,7))

bins = np.linspace(0.305, 1.0001, 14)

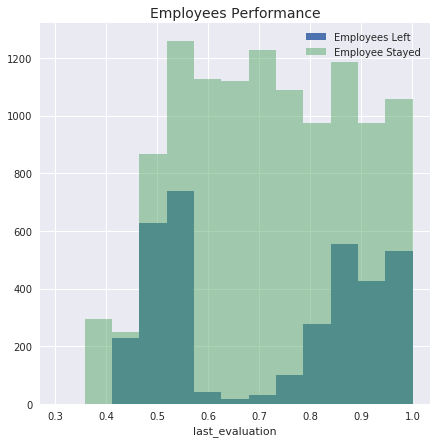
plt.hist(data[data['left']==1]['last\_evaluation'], bins=bins, alpha=1, label='Employees Left')

plt.hist(data[data['left']==0]['last\_evaluation'], bins=bins, alpha = 0.5, label = 'Employee Stayed')

plt.title('Employees Performance', fontsize=14)

plt.xlabel('last\_evaluation')

plt.legend(loc='best');



It is possible to see that 98% of employees with few projects that left also have poor performance.

And 95% of the employees with 5 or more projects that left the company had the highest performance.

3 or 4 are the best number of projects.

In [28]:

poor\_performance\_left = data[(data.last\_evaluation <= 0.62) & (data.number\_project == 2) & (data.left == 1)]

print('poor\_performance\_left:',len(poor\_performance\_left))

poor\_performance\_stayed = data[(data.last\_evaluation > 0.62) & (data.number\_project == 2) & (data.left == 1)]

print('poor\_performance\_stayed:',len(poor\_performance\_stayed))

print('**\n**')

high\_performance\_left= data[(data.last\_evaluation <= 0.62) & (data.number\_project >=5) & (data.left == 1)]

high\_performance\_stayed= data[(data.last\_evaluation > 0.8) & (data.number\_project >=5) & (data.left == 0)]

print('high\_performance\_left:',len(high\_performance\_left))

print('high\_performance\_stayed', len(high\_performance\_stayed))

plt.figure(figsize =(7,5))

bins = np.linspace(1.5,7.5, 7)

plt.hist(data[data['left']==1]['number\_project'], bins=bins, alpha=1, label='Employees Left')

plt.hist(data[data['left']==0]['number\_project'], bins=bins, alpha = 0.5, label = 'Employee Stayed')

plt.title('Number of projects', fontsize=14)

plt.xlabel('number\_ projects')

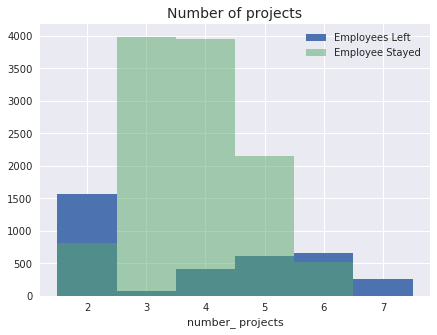
plt.legend(loc='best');

poor\_performance\_left: 1531

poor\_performance\_stayed: 36

high\_performance\_left: 47

high\_performance\_stayed 889



### Working hours

Again, there are 2 groups of employees. A group that works fewer hours and another that works more hours compared to the average hours worked.

In [29]:

plt.figure(figsize =(7,5))

bins = np.linspace(80,315, 15)

plt.hist(data[data['left']==1]['average\_montly\_hours'], bins=bins, alpha=1, label='Employees Left')

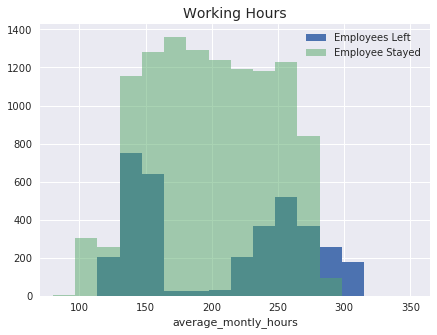
plt.hist(data[data['left']==0]['average\_montly\_hours'], bins=bins, alpha = 0.5, label = 'Employee Stayed')

plt.title('Working Hours', fontsize=14)

plt.xlabel('average\_montly\_hours')

plt.xlim((70,365))

plt.legend(loc='best');



Clearly is possible to see that the employees with 6 projects or more, work on average 20% more hours.

In [30]:

groupby\_number\_projects = data.groupby('number\_project').mean()

groupby\_number\_projects = groupby\_number\_projects['average\_montly\_hours']

print(groupby\_number\_projects)

plt.figure(figsize=(7,5))

groupby\_number\_projects.plot();

number\_project

2 160.342546

3 197.507522

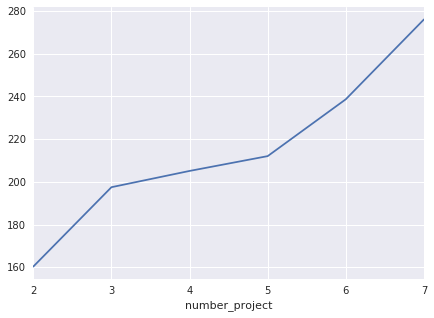
4 205.122108

5 212.061572

6 238.694208

7 276.078125

Name: average\_montly\_hours, dtype: float64



With the information above the employees that left the company are grouped as:

* Employees with 2 projects and worked less than the average of the company.
* Employees with 5 or more projects that worked at least 20% more than the average.

In [31]:

work\_less\_hours\_left = data[(data.average\_montly\_hours < 200) & (data.number\_project == 2) & (data.left == 1)]

print('work\_less\_hours\_left:',len(work\_less\_hours\_left))

work\_more\_hours\_left = data[(data.average\_montly\_hours > 240) & (data.number\_project >=5 ) & (data.left == 1)]

print('work\_more\_hours\_left:',len(work\_more\_hours\_left))

*#<p><font color="red">Aqui você fala sobre a relação entre horas de trabalho e quantidade de projetos, mas isso não é exibido no gráfico</font></p>*

work\_less\_hours\_left: 1535

work\_more\_hours\_left: 1225

### Satisfaction Level

It is possible to see 3 interesting peaks in the satisfaction levels of the employees that left the company.

* We have a peak of employees who are totally disappointed.
* Another peak at 0.4, representing another group with the satisfaction level below the average.
* And another amount in the range 0.7 and 0.9, with employees that left, although the high satisfaction.

In [32]:

plt.figure(figsize =(7,5))

bins = np.linspace(0.006,1.000, 15)

plt.hist(data[data['left']==1]['satisfaction\_level'], bins=bins, alpha=1, label='Employees Left')

plt.hist(data[data['left']==0]['satisfaction\_level'], bins=bins, alpha = 0.5, label = 'Employee Stayed')

plt.title('Employees Satisfaction', fontsize=14)

plt.xlabel('satisfaction\_level')

plt.xlim((0,1.05))

plt.legend(loc='best');



#### Average satisfaction for years in the company

In [33]:

groupby\_time\_spend = data.groupby('time\_spend\_company').mean()

groupby\_time\_spend['satisfaction\_level']

Out[33]:

time\_spend\_company

2 0.697078

3 0.626314

4 0.467517

5 0.610305

6 0.603440

7 0.635957

8 0.665062

10 0.655327

Name: satisfaction\_level, dtype: float64

#### When the employees becames unsatisfayed?

In next results it is clear the drop in satisfaction when employees are working on 6 or more projects.

In [34]:

sns.set()

sns.set\_context("talk")

ax = sns.factorplot(x="number\_project", y="satisfaction\_level", col="time\_spend\_company",col\_wrap=4, size=3, color='blue',sharex=**False**, data=data)

ax.set\_xlabels('Number of Projects');



Let´s see why the most valuable employees tend to leave.

From the employees that left with high performance, 4 or more years in the company and working on 5 or more project had:

* Low satisfaction level,
* Worked more hours,
* Haven´t been promoted in the last five years.

In [35]:

func\_living = data[(data.last\_evaluation >= 0.70) | (data.time\_spend\_company >=4) | (data.number\_project >= 5)]

corr2 = func\_living.corr()

sns.set(style='white')

mask = np.zeros\_like(corr2, dtype=np.bool)

mask[np.triu\_indices\_from(mask)] = **True**

*# Insert the graphic*

f, ax = plt.subplots(figsize=(13,8))

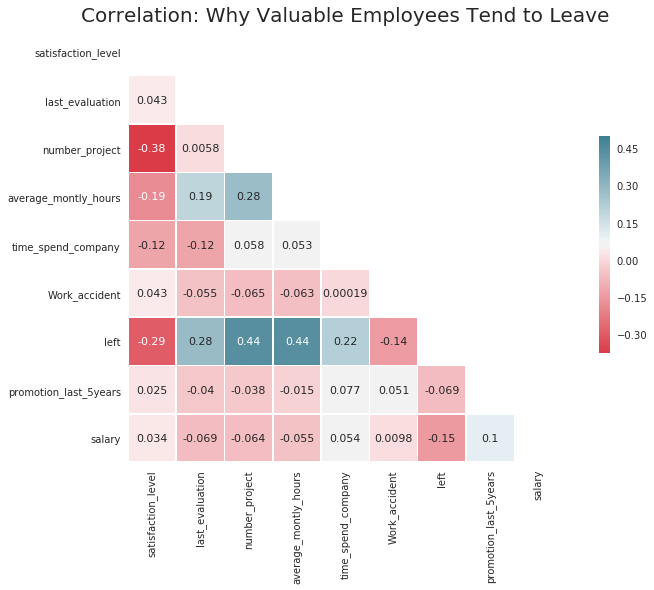
cmap = sns.diverging\_palette(10,220, as\_cmap=**True**)

*#Draw heat map mask*

ax = sns.heatmap(corr2, mask=mask, cmap=cmap, vmax= .5, annot=**True**, annot\_kws= {'size':11}, square=**True**, xticklabels=**True**, yticklabels=**True**, linewidths=.5,

cbar\_kws={'shrink': .5}, ax=ax)

ax.set\_title('Correlation: Why Valuable Employees Tend to Leave', fontsize=20);



### **Summary of the Exploratory Data Analysis**

* It is a relatively young company, on average, employees have 3 or 4 years in the company and the oldest employees are working 10 years.
* The biggest difference in the salary from who stayed and those who left, was found in the management department, in the others departments although the salaries of who stayed be higher in average, it is not a big difference.
* The number of employees that had a work accident is about 14%, of which only 169 employees left the company, so don't seem to have a correlation with the employees leaving.
* In five years only 2% of the employees were promoted. Is possible that many employees get unmotivated and start planning to leave.
* Employees with 7 or longer in the company didn't left. Employees with 5 years have more chances to leaving.
* There are 2 distinct groups of employees performance that left. A group with poor performance with 2 projects and others with high performance with 5 or more projects. It is not necessary retain all the employees, the focus is on keeping employees with high performance.
* The employees with 4 years in the company have the lowest average satisfaction level of all the company with (0.47).
* The satisfaction drops when the employees are working in 5 or more projects. A number of 3 or 4 projects seems to be ideal independent of the time spend in the company.
* The employees with 5 or more projects that left also worked at least 20% more hours than the average of the company.
* The satisfaction level of the employees that left is grouped in totally disappointed, below the average satisfaction and satisfied.

**PROPOSED SYSTEM TO CATER FOR GENDER DIVERSITY**

# Human Resources Diversity Analysis

This is a Descriptive Analytics Project Here they are:

1. Are there areas of the company where pay is not equitable?
2. What is the overall diversity profile of the organization?

We will be analyzing each matter with regards to race, gender and age.

## Importing what's important

These are standard libraries for EDA. Here we go:

In [1]:

**import** **matplotlib.pyplot** **as** **plt**

**import** **seaborn** **as** **sns**

**import** **pandas** **as** **pd**

**import** **numpy** **as** **np**

%**matplotlib** inline

## Exploring the data

### First Look

Let's load the dataset and take a peek

In [2]:

hr\_data = pd.read\_csv('../input/human-resources-data-set/HRDataset\_v13.csv')

hr\_data.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 401 entries, 0 to 400

Data columns (total 35 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Employee\_Name 310 non-null object

1 EmpID 310 non-null float64

2 MarriedID 310 non-null float64

3 MaritalStatusID 310 non-null float64

4 GenderID 310 non-null float64

5 EmpStatusID 310 non-null float64

6 DeptID 310 non-null float64

7 PerfScoreID 310 non-null float64

8 FromDiversityJobFairID 310 non-null float64

9 PayRate 310 non-null float64

10 Termd 310 non-null float64

11 PositionID 310 non-null float64

12 Position 310 non-null object

13 State 310 non-null object

14 Zip 310 non-null float64

15 DOB 310 non-null object

16 Sex 310 non-null object

17 MaritalDesc 310 non-null object

18 CitizenDesc 310 non-null object

19 HispanicLatino 310 non-null object

20 RaceDesc 310 non-null object

21 DateofHire 310 non-null object

22 DateofTermination 103 non-null object

23 TermReason 309 non-null object

24 EmploymentStatus 310 non-null object

25 Department 310 non-null object

26 ManagerName 310 non-null object

27 ManagerID 302 non-null float64

28 RecruitmentSource 310 non-null object

29 PerformanceScore 310 non-null object

30 EngagementSurvey 310 non-null float64

31 EmpSatisfaction 310 non-null float64

32 SpecialProjectsCount 310 non-null float64

33 LastPerformanceReview\_Date 207 non-null object

34 DaysLateLast30 207 non-null float64

dtypes: float64(17), object(18)

memory usage: 109.8+ KB

In [3]:

*# Lots of empty rows. Some cleaning now:*

hr\_data.dropna(how='all', inplace=**True**)

### Racial Diversity

In [4]:

*# A first look at the racial diversity situation - both in absolute and percentual figures:*

hr\_data.RaceDesc.name = 'Racial group'

display(hr\_data.RaceDesc.value\_counts(),

hr\_data.RaceDesc.value\_counts(normalize=**True**) \* 100)

White 193

Black or African American 57

Asian 34

Two or more races 18

American Indian or Alaska Native 4

Hispanic 4

Name: Racial group, dtype: int64

White 62.258065

Black or African American 18.387097

Asian 10.967742

Two or more races 5.806452

American Indian or Alaska Native 1.290323

Hispanic 1.290323

Name: Racial group, dtype: float64

According to 2018 data by the U.S Bureau of Labor Statistics, the country's labor force is made of 78% Whites, 13% Blacks and 6% Asians (<https://www.bls.gov/opub/reports/race-and-ethnicity/2018/home.htm>). Taking that as a baseline, we can say that the company is diverse from a racial standpoint.

**Let's see which recruitment sources hire the most non-white employees:**

In [5]:

hr\_data['Non-white'] = (hr\_data['RaceDesc'] != 'White')

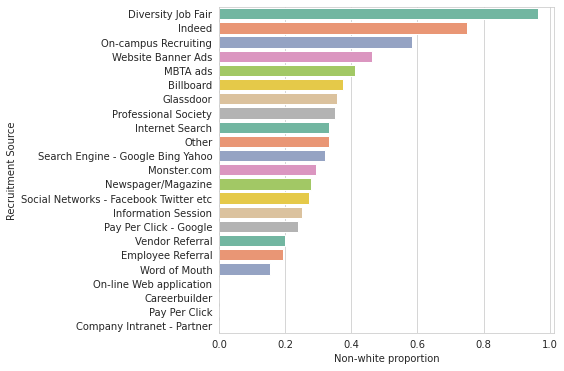
fig = plt.figure(figsize=(6, 6)), sns.set\_style('whitegrid')

*# Ordering for better visualization*

ordered\_nw = hr\_data.groupby('RecruitmentSource')['Non-white'].mean().reset\_index().sort\_values('Non-white', ascending=**False**)

ax = sns.barplot(data=ordered\_nw, y='RecruitmentSource', x='Non-white', palette='Set2')

ax.set\_xlabel('Non-white proportion'), ax.set\_ylabel('Recruitment Source');



Unsurprisingly, **Diversity Job Fair** plays a crucial role in promoting racial diversity. **Indeed and On-campus Recruiting** can also be lauded for bringing non-white employees more often than not.

On the other hand, **Pay Per Click, On-line Web application, Careerbuilder and Company Intranet** have no contribution to racial diversity at all.

Of course, **respect to diversity includes egalitarian compensation, regardless of skin color.** Let's take a look at that now.

In [6]:

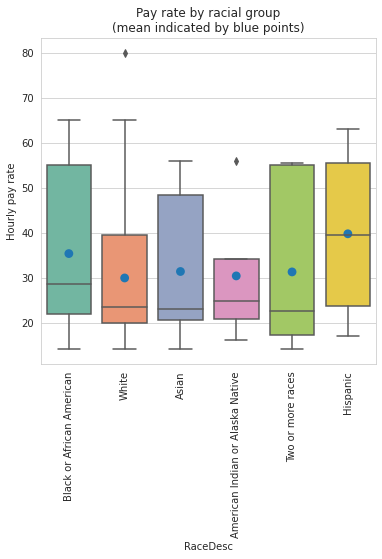
plt.figure(figsize=(6, 6))

ax = sns.boxplot(data=hr\_data, x='RaceDesc', y='PayRate', palette='Set2')

ax.set\_xticklabels(hr\_data.RaceDesc.unique(), rotation=90)

sns.pointplot(data=hr\_data, x='RaceDesc', y='PayRate', join=**False**, ci=**None**, ax=ax);

ax.set\_ylabel('Hourly pay rate'); ax.set\_title('Pay rate by racial group**\n**(mean indicated by blue points)');



There are a few differences in median and mean payment across races. For instance, Hispanics are placed a bit higher. This does not, however, seem particularly problematic. Let's look at each department.

In [7]:

g = sns.catplot(data=hr\_data, x='RaceDesc', col\_wrap=2,

y='PayRate', col='Department', kind='box', palette='Set2')

g.set\_xticklabels(hr\_data.RaceDesc.unique(), rotation=90)

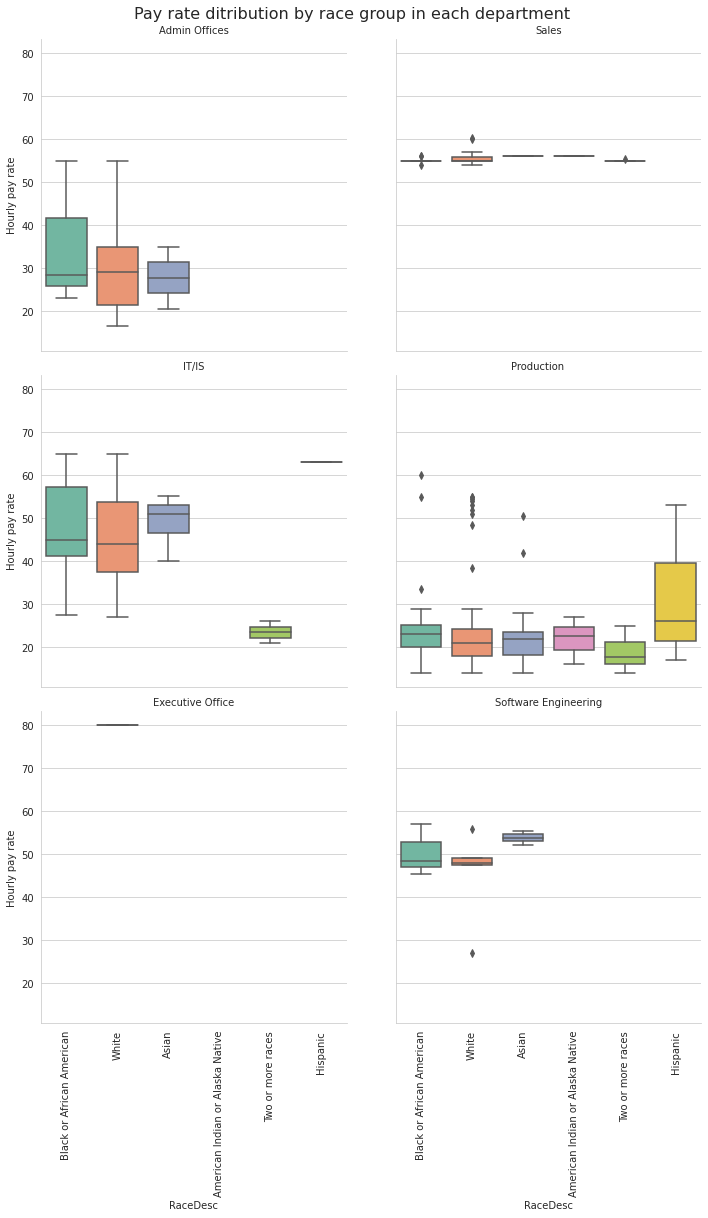
g.set\_ylabels('Hourly pay rate')

g.fig.suptitle(

'Pay rate ditribution by race group in each department', fontsize=16)

g.set\_titles('**{col\_name}**')

plt.subplots\_adjust(top=0.95)



An in-depth look at each department shows that payment is generally egalitarian. **The exception is IT/IS, wherein those of two or more races have a diminished pay rate.** Why is that? Are these workers in low-paid positions?

In [8]:

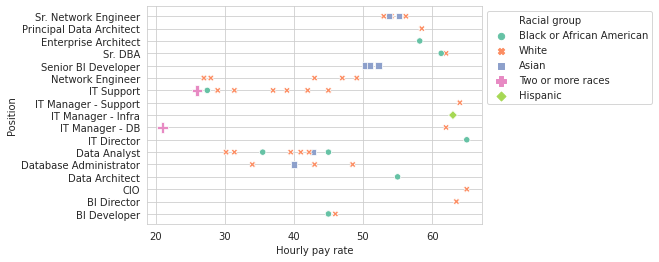
*# Removing whitespace in "Data Analyst " (Thanks for noticing, Jordan!)*

hr\_data['Position'] = hr\_data['Position'].str.strip()

ax = sns.scatterplot('PayRate', 'Position', 'RaceDesc', data=hr\_data.loc[hr\_data.Department=='IT/IS'],

palette='Set2', style='RaceDesc',size='RaceDesc', sizes=[40, 40, 40, 120, 40]);

ax.legend(bbox\_to\_anchor=(1, 1)).texts[0].set\_text('Racial group'); ax.set\_xlabel('Hourly pay rate');



The plot reveals that employees of this racial group are actually being paid lower than colleagues in the same position. This could be the result of discriminatory practices. Let's investigate further to see if there are any reasonable causes for the wage gap:

In [9]:

*# Filtering only rows that contain "two or more races" workers*

position\_rows = hr\_data.Position.isin(['IT Support', 'IT Manager - DB'])

perf\_indicators = ['RaceDesc','Position', 'PerformanceScore', 'SpecialProjectsCount', 'DaysLateLast30','EngagementSurvey']

it\_is\_lookup = hr\_data.loc[position\_rows, perf\_indicators].sort\_values(['Position','RaceDesc']).set\_index('RaceDesc')

it\_is\_lookup

Out[9]:

|  | **Position** | **PerformanceScore** | **SpecialProjectsCount** | **DaysLateLast30** | **EngagementSurvey** |
| --- | --- | --- | --- | --- | --- |
| **RaceDesc** |  |  |  |  |  |
| **Two or more races** | IT Manager - DB | Fully Meets | 6.0 | NaN | 2.51 |
| **White** | IT Manager - DB | Fully Meets | 7.0 | 0.0 | 2.96 |
| **Black or African American** | IT Support | Fully Meets | 5.0 | 0.0 | 4.30 |
| **Two or more races** | IT Support | Exceeds | 5.0 | 0.0 | 4.64 |
| **White** | IT Support | Fully Meets | 6.0 | 0.0 | 2.55 |
| **White** | IT Support | Fully Meets | 5.0 | 0.0 | 1.21 |
| **White** | IT Support | Fully Meets | 6.0 | 0.0 | 1.84 |
| **White** | IT Support | Fully Meets | 7.0 | 0.0 | 2.21 |
| **White** | IT Support | Fully Meets | 6.0 | 0.0 | 4.11 |
| **White** | IT Support | Fully Meets | 5.0 | 0.0 | 4.61 |

The perfomance of two or more races employees is rather levelled with that of colleagues in the same position, yet their pay rates are the lower than any other. **As far as the dataset goes, this points to the existence of discrimination.**

**DISCLAIMER: here and throughout the end of this analysis, there is an observation that distinguishes itself from the others: at Executive Office, we find the President & CEO of the company, who is naturally the most well-paid. Take a look at her:**

In [10]:

hr\_data.loc[hr\_data['Department'] == 'Executive Office']

Out[10]:

|  | **Employee\_Name** | **EmpID** | **MarriedID** | **MaritalStatusID** | **GenderID** | **EmpStatusID** | **DeptID** | **PerfScoreID** | **FromDiversityJobFairID** | **PayRate** | **...** | **ManagerName** | **ManagerID** | **RecruitmentSource** | **PerformanceScore** | **EngagementSurvey** | **EmpSatisfaction** | **SpecialProjectsCount** | **LastPerformanceReview\_Date** | **DaysLateLast30** | **Non-white** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **73** | King, Janet | 1.001495e+09 | 1.0 | 1.0 | 0.0 | 1.0 | 2.0 | 3.0 | 0.0 | 80.0 | ... | Board of Directors | 9.0 | Pay Per Click - Google | Fully Meets | 4.83 | 3.0 | 0.0 | 1/17/2019 | 0.0 | False |

1 rows × 36 columns

### Gender Equality

In [11]:

*# For clarity:*

hr\_data.replace({'Sex': {'F': 'Female', 'M ': 'Male'}}, inplace=**True**)

hr\_data.Sex.name = 'Gender'

*# Now, to an overview in gender distribution:*

print(hr\_data.Sex.value\_counts(),'**\n\n**', (hr\_data.Sex.value\_counts(normalize=**True**) \* 100), sep='')

Female 177

Male 133

Name: Gender, dtype: int64

Female 57.096774

Male 42.903226

Name: Gender, dtype: float64

The numbers are balanced, and **most employees are female.**

Taking a look at each department now:

In [12]:

*# By department*

g = sns.catplot(data=hr\_data, x='Sex', col='Department',

col\_wrap=2, palette='Spectral', kind='count')

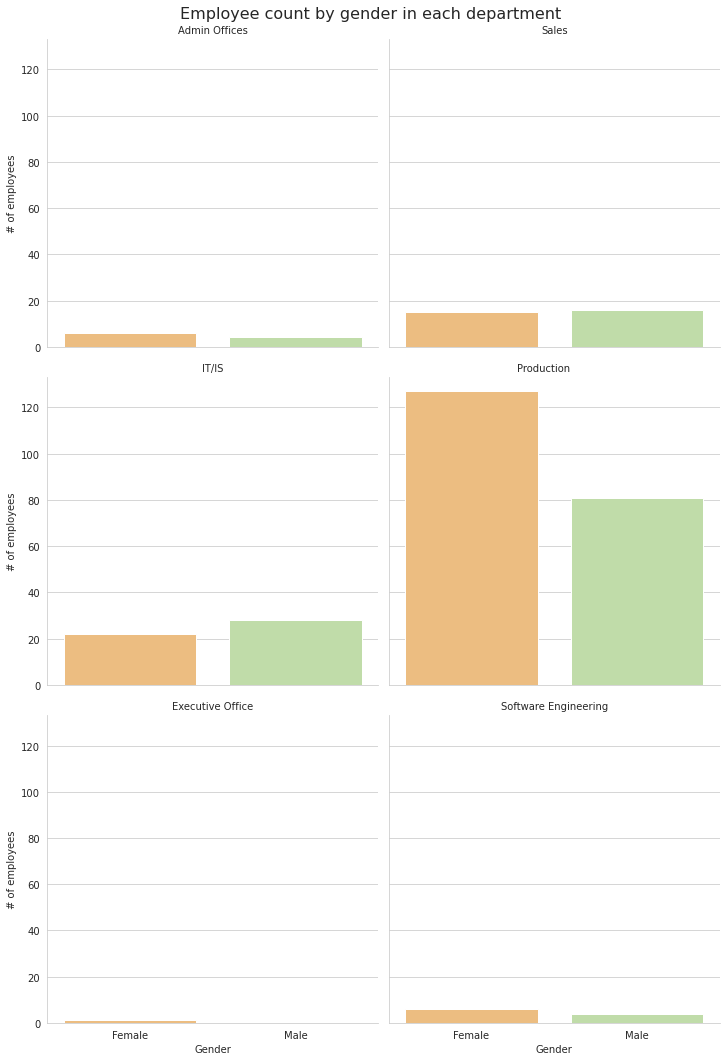
g.set\_xlabels('Gender'), g.set\_ylabels('# of employees')

g.fig.suptitle(

'Employee count by gender in each department', fontsize=16)

g.set\_titles('**{col\_name}**')

plt.subplots\_adjust(top=0.95)



We can see that the numeric difference stems from the Production department, where women outnumber men by 50%.

**But are female employees getting equal pay?**

In [13]:

GenderPay = hr\_data.groupby('Sex')[['PayRate']]

display(GenderPay.agg(['mean', 'median']))

|  | **PayRate** | |
| --- | --- | --- |
|  | **mean** | **median** |
| **Sex** |  |  |
| **Female** | 29.472147 | 24.0 |
| **Male** | 33.697143 | 26.0 |

**That's a no.** Women's income is lower in general.

How does that distribute across departments?

In [14]:

*# Removing whitespace in 'Production':*

hr\_data['Department'] = hr\_data['Department'].str.strip()

plt.figure(figsize=(8, 6))

ax = sns.pointplot(data=hr\_data, x='Department', y='PayRate',

hue='Sex', palette='Spectral' , join=**False**)

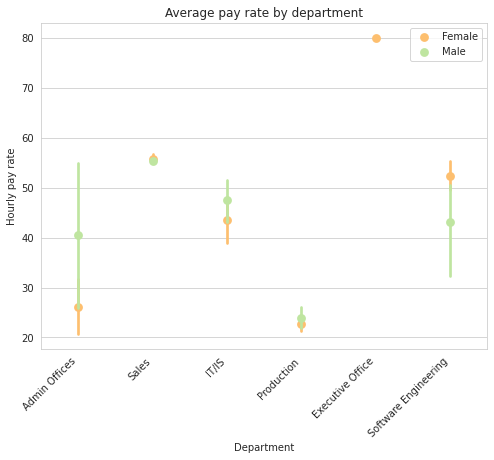
ax.set\_xticklabels(hr\_data.Department.unique(), rotation=45,

horizontalalignment='right');

ax.legend()

ax.set\_title("Average pay rate by department"), ax.set\_ylabel(

'Hourly pay rate');



At the plot, we identify some divergences in average pay by gender. Women's average pay is quite higher at Software Engineering department. However, it is important to note that only a handful of employees work there, in such a way that it doesn't affect the overall statistics so much.

On the other hand, Production plays a big part in broadening the wage gap because:

1. the department has the lowest pay rate;
2. this decreases female mean compensation by a lot, **since it contains the largest number of workers, most of them being women**;
3. considering the employee amount, even though the inequality within Production itself isn't great, it ends up making a significant impact.

**We can see a substantial wage gap at Admin Offices**; there, women's average income is much lower than men's.

It demands some examining:

In [15]:

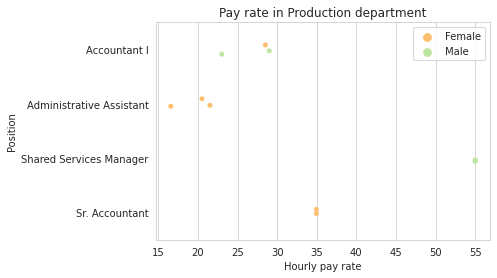
ax = sns.stripplot('PayRate', 'Position', hue='Sex',

data=hr\_data[hr\_data.Department == 'Admin Offices'],

palette='Spectral')

ax.set\_xlabel('Hourly pay rate'), ax.set\_title('Pay rate in Production department')

ax.legend();



An in-depth analysis reveals that the wage disparity at Admin Offices, the largest throughout departments, **comes from the presence of two men working in a high-paid position (shared services manager), contrasted with three women employed as administrative assistant.**

Among those whose title is Accountant I, the female employee's pay rate is only marginally lower than that of a male colleague. We can also see that only women work as Sr. Accountants, **which shows that females aren't necessarily kept away from higher positions in this department**.

### Age Diversity

Now, you might have noted that there's no age info in the dataset. Well, at least not yet.

We can obtain it through the date of birth (DOB) column

In [16]:

*# We can make our own age column.*

*# But first, let's convert 'DOB' to datetime format, with a bit of chaining to keep just the date*

**from** **dateutil.relativedelta** **import** relativedelta

hr\_data['DOB'] = pd.to\_datetime(hr\_data['DOB']).dt.date.astype('datetime64')

*# No employees were born after year 2000, so DOBs like 2068 should have 100 years removed:*

hr\_data.loc[hr\_data.DOB.dt.year > 2000, 'DOB'] -= pd.DateOffset(years=100)

*# Now, to getting the age:*

hr\_data['Age'] = pd.Series(dtype='int')

**for** ind, date **in** hr\_data.DOB.iteritems():

hr\_data.loc[ind, 'Age'] = relativedelta(

pd.to\_datetime('today'), date).years

hr\_data.Age = hr\_data.Age.astype('int64')

*#Some summary statistics*

hr\_data['Age'].apply(['min', 'median', 'max'])

Out[16]:

min 27.0

median 39.0

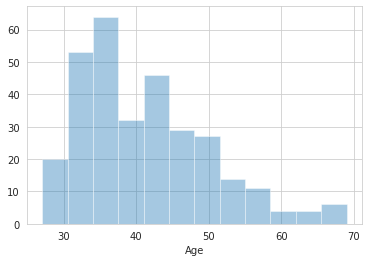
max 69.0

Name: Age, dtype: float64

Now we have the employees' ages stored in a column. They range from 27 to 69; the median is 39. Let's visualize how the number of workers is distributed by age, and also take a look at percentages by age group:

In [17]:

ax = sns.distplot(hr\_data['Age'], kde=**False**)



In [18]:

agegroups= pd.cut(hr\_data.Age, [25,40,55,70])

agegroups.name = 'Age group'

agegroups.value\_counts(normalize=**True**) \*100

Out[18]:

(25, 40] 54.516129

(40, 55] 39.032258

(55, 70] 6.451613

Name: Age group, dtype: float64

The staff's ages are mainly at early to mid-30s, also counting high at early 40s. **The number of employees is substantially lower for ages 55 and over, amounting only to no more than a few dozens, or 6.4%.**

Considering how low the numbers go when it comes to older workers, it's important that we inquire how age diversity is promoted through recruiting.

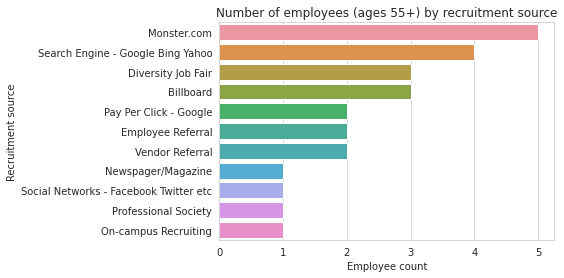
In [19]:

ordered\_55 = hr\_data.loc[hr\_data.Age >= 55, 'RecruitmentSource'].value\_counts()

ax = sns.barplot(x=ordered\_55.values,y=ordered\_55.index)

ax.set\_ylabel('Recruitment source'), ax.set\_xlabel('Employee count'), ax.set\_xticks(range(6))

ax.set\_title('Number of employees (ages 55+) by recruitment source');



**Monster.com** scores the highest,hiring five mature employees, followed by **search engines**, with four. The numbers are really low, in such a way that no source is currently bringing a reasonable amount of older people to the company.

**Are salaries subject to ageism at Dental Magic?** It is necessary to take a look at how mature workers are being payed.

In [20]:

g = sns.relplot('Age', 'PayRate', agegroups, hue\_order=agegroups.cat.categories, col='Department',

col\_wrap=2, data=hr\_data, palette='PuRd')

g.set\_xlabels('Age'), g.set\_ylabels('Hourly pay rate')

g.fig.suptitle('Age x Pay rate across departments', fontsize=16)

g.set\_titles('**{col\_name}**')

plt.subplots\_adjust(top=0.95)



We saw before that the employee count goes down for ages 50 and over. In the Production department, we can see that **salaries are also diminished**: these workers have a pay rate below 35, whereas some younger ones get as high as $60/hour. This calls for a more detailed analysis of this specific division.

Although the plot is focused on salary, it also reveals that both Admin Offices and Software Engineering barely have any workers over 40.

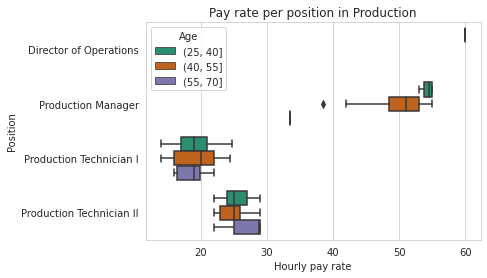
In [21]:

agegroups= pd.cut(hr\_data.Age, [25,40,55,70])

ax = sns.boxplot('PayRate', 'Position', agegroups,

data=hr\_data.loc[hr\_data.Department=='Production'], palette='Dark2') ;

ax.set\_xlabel('Hourly pay rate'); ax.set\_title('Pay rate per position in Production');



The graph clarifies that it is generally not the case that older workers are getting a lower salary than their position's average. **Matures' pay rates are limited because almost all of them work in lesser functions (Production Technician I and II).**

**In Software Engineering, there is an outlier: a worker just over 50 years old, whose pay rate is much lower than that of younger colleagues**. We will now take a closer look at the department to shed light on the matter.

In [22]:

*# The department:*

soft\_engs = hr\_data.Department == 'Software Engineering'

*# Columns for analysis:*

relevant\_info = ['Employee\_Name', 'Age', 'Position', 'PayRate',

'PerformanceScore', 'EmploymentStatus', 'EngagementSurvey']

*# Now, we select them and look at some relevant columns:*

hr\_data.loc[soft\_engs, relevant\_info].sort\_values('Age', ascending=**False**).set\_index('Age')

Out[22]:

|  | **Employee\_Name** | **Position** | **PayRate** | **PerformanceScore** | **EmploymentStatus** | **EngagementSurvey** |
| --- | --- | --- | --- | --- | --- | --- |
| **Age** |  |  |  |  |  |  |
| **53** | Sweetwater, Alex | Software Engineering Manager | 27.00 | Fully Meets | Active | 3.84 |
| **41** | Andreola, Colby | Software Engineer | 47.60 | Fully Meets | Active | 3.04 |
| **41** | Del Bosque, Keyla | Software Engineer | 57.12 | Fully Meets | Active | 3.79 |
| **41** | Patronick, Luke | Software Engineer | 52.25 | Exceeds | Voluntarily Terminated | 1.10 |
| **37** | Szabo, Andrew | Software Engineer | 48.00 | Exceeds | Active | 2.61 |
| **37** | True, Edward | Software Engineer | 45.42 | Fully Meets | Voluntarily Terminated | 1.11 |
| **33** | Carabbio, Judith | Software Engineer | 56.00 | Fully Meets | Active | 4.96 |
| **33** | Exantus, Susan | Software Engineer | 48.50 | Needs Improvement | Terminated for Cause | 2.55 |
| **33** | Saada, Adell | Software Engineer | 49.25 | Fully Meets | Active | 1.74 |
| **32** | Martin, Sandra | Software Engineer | 55.51 | Fully Meets | Active | 1.53 |

We now see that the outlier is Alex Sweetwater.

He is the manager of the department, but his pay is registered as the lowest one. It is highly unlinkely that his pay rate is lower than his subordinates'. Also, the data doesn't point to bad performance. **It might be sensible to assume that this piece of information is actually incorrect, possibly due to input error.**

## Sharing findings and insights

### 1. What are our best recruiting sources if we want to ensure a diverse organization?

It's important that every organization strives to make sure their recruitment practices aren't affected by prejudices or biases. This matter should be addressed with regards to which groups are rejected the most when searching for jobs. **Discrimination is an issue that constantly hits non-white people, women and elders.**

Throughout the analysis of all features, Diversity Job Fair proved to be vital in making Dental Magic more plural. It should only be encouraged and expanded.

Regarding race, more than half of the people hired via Indeed and on-campus recruiting are from underrepresented groups. Contrastingly, some sources need to undergo further scrutiny as to why they only bring white employees.

When it comes to age diversity, the organization is far behind, and no recruitment source is distinctly efficient.

### 2. Are there areas of the company where pay is not equitable?

A deep analysis has showed some wage gaps inside departments:

1. **At IT/IS, people of two or more races are paid significantly less than other workers in the same job position**. None of the data show a cause for that.
2. **Women's income is lower overall.** The gap stems mainly from two sources:
   * in Production, the least well-paid department, females outnumber males by some extent, which results in a greater impact in women's overall salary. Their average income is also slightly inferior in the department;
   * in Admin Offices, the wage gap between genders is substantial, though it is hard to tell if the distribution of functions is discriminatory.
3. **Most workers aged 50+ in Production work at lower-paid positions.**

### 3. What is the overall diversity profile of the organization?

**Race**

While more than half of the workforce is made of white people, we've seen before that the jobs are more well distributed along underrepresented groups comparatively to official statistics in a national level.

Hispanics get a slightly higher pay rate in average, whereas American Indian or Alaska Natives perform lower on that variable. Considering that these groups have only four members each, it is premature to conclude the divergence is caused by any discriminatory treatment.

**Gender**

The workforce is predominantly female. Additionally, Dental Magic's CEO is a woman, and some more can be found in other high positions. These are positive, distinctive traits in a world that favors hiring male workers for most roles, especially leading ones.

A potential highlight is how the company hires many women to work in Production, a department where labor is often manual.

However, the company still faces some income inequality related to gender. The issue should be further investigated and dealt with.

**Age**

Only 6.4% of the workers are 55 or older. This is certainly a diversity issue, specially looking at how 2018 data from the US BLS shows that 23.1% of the workforce in the country is in that age group (<https://www.bls.gov/emp/graphics/2019/labor-force-share-by-age-group.htm>).

**The matter is specially precarious in the Software Engineering department and Admin Offices, that together count no more than a handful employees over 40.**

Sales, on the other hand, performs really well, having a good amount of well-paid elderly workers.

### Final thoughts

The biggest issue is the lack of age diversity, which is not truly promoted by any recruitment source. The organization should review its hiring practices to remove any potential bias.