Quentin JONNEAUX
Student Number : R00274704
DATA 8001 - DATA SCIENCE AND ANALYTICS

'DENTAL MAGIC - DATA SCIENCE AND ANALYTICS PROJECT'

MTU
Ollscoil Teicneolaíochta na Mumhan
Munster Technological University

Table of content

Site Assessment

Business description

Dental Magic is a US company specialised in the production of Dentistry products, having its production based in Massachusetts. The company has spread its sales workforce throughout USA, with a repartition between the east and west coast. Except sales, most departments seem centralized in Massachusetts (Admin, Executive, IT, Software Engineering and Production).

The company is seeking consultancy on their HR data to understand several aspects of their workforce including:
- Relationship between performance and person reporting to
- Diversity of the organization
- Best recruiting source to ensure diversity
- Predicting who is going to terminate
- Pay inequities within same area

We will use Clarke's Maturity questionnaire to assess the site.

Business Strategy

BIG DATA STRATEGY

The company seems to understand the value of Big Data. It is highlighted by the fact that consultancy was asked, specifically on their workforce to understand how to pilot it. Business and IT leaders are questioning the need to look at the data hosted and is looking to start an Analytics project, with a budget of 50,000 €. The company is also well-equipped in terms of people that could be dedicated to business intelligence and analytics reporting. The company can currently count on:
- 4 Bi Developers
- 3 Senior BI Developers
- 1 BI director
- 6 Data Analysts
- 1 Data Architect
- 2 Database Administrators
- 1 Senior Database Administrators

The company has the capacity to launch Analytics projects but strategy needs adjustments.

BIG DATA USAGE

The company does not seem however to be confident to use anything else that analytical base tables to pilot business. It seems some data is scanned for some information but most data seems to be entered manually. Some designed tools are dedicated to generate regular reports but rarely dedicates audits. One reason is that sales department is widespread with US states, which would result in traveling a lot. In some rare instances, the production is audited but always at significant cost for the company.

However, the company started cataloging their data for analytical purposes and decided the HR dataset was suitable to start their first analytics project. Since HR contains information on employees, the data has also been assessed for any compliance issue with GDPR and CCPA. Since the company is new to Analytics, compliance against Eu's ethics guidelines for AI has not been assessed yet.

IDEA FOR INCREASE

- AI Adoption: based on a McKinsey study, companies adopting AI in HR, 55% acknowledged a decreased in average cost (49% between 0 and 19%). AI can be leveraged to save money on HR. Gen AI especially can be used to explore use cases.

- Implementation Audit automation: with Data Governance becoming more and more focused, companies will need to ensure compliance with GDPR, CCPA and AI acts. Automating audits seems to be a must to comply with regulations

Process and Skills

ORGANISATIONAL MATURITY

Dental Magic recently understand the value of Big Data so the company started to train some Data Analysts to onboard on the project. The bright side of their workforce, with the numbers of analytics specialists named above is that they have a number of tools and resources with data visualisation, in terms of predictive and forecasting. Analysts and BI team are well-trained on Power BI and Tableau and they are regularly fed with Data to pilot the business. In addition of Tableau and Power BI, the team mostly used Python for scientific programming, due its versatility as other departments rely on this language (Automation for production, IT and software development) The company is also looking for ways standardise analytics over the site.

ANALYTICS ARCHITECTURE

Their BI Toolset is composed of leading BI tools such as Power BI, Tableau and SAP and analysts are trained on their predictive tools. The company is looking to augment their toolset with Data Science skills. MySQL is used as Relational Database Management System to provide Analytical Base table, for its ease of integration. In terms of ETL capabilities, the company uses AWS Data pipeline for its inexpensive aspect and fault-tolerance feature (as the company is starting projects on analytics). Despite not being open-source, the tool is easy to use and can scale to cloud and integrates with our AWS services if needed. Due to the company presence in different states and recent understanding on big data, the performance could be improved in terms of response time.

IDEA FOR INCREASE

- Train some Analyst using R to leverage for Data Manipulation aspects (quicker than Python to use excel sheets), custom visualization (dplyr, ggplot) and statistical analysis (simulation, hypothesis testing)
- Explore predictive capabilities of BI tools with python
- Consider NoSQL approach such as MongoDB to start gathering unstructured data (MongoDB is easy to integrate, aligning with the ease of use aspect of most tools) and earn availability and performance
- Start to build a Data Science team with at least one fully dedicated Data Scientist, a number of tools and a portfolio of project to address

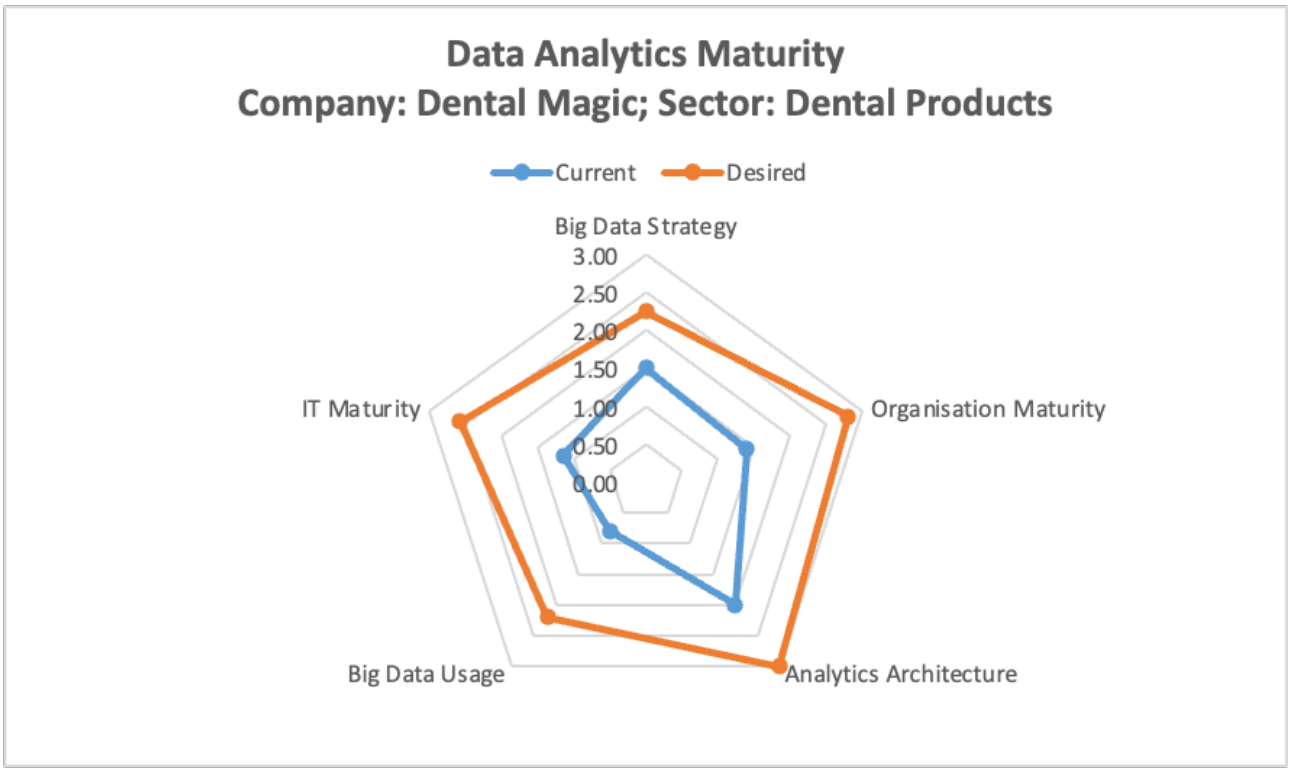
Tech Skills

IT MATURITY

Dental Magic seems to have bought the idea of big data but it seems the company has a strategy of working in silos, meaning each department seems to have a different strategy to align data with each other. They are looking to build an operated Big Data model but the strategies need to be reshape to reflect the site and not a combination of silos before defining an operational model. Each service adopted Big Data and most services have standardized systems of reporting. Some departments are using BI tools to gather insights but most data seems to be stored on premises. The company is transitioning to big data and are looking for projects that could contribute to this transition instead having only short-term deliverables.

IDEA FOR INCREASE

- Break the silos to align the Big Data strategy to the entire site
- Standardized system may remain but consider adding a layer to translate a standard to another, to make data profitable for other services
- Consider Cloud solutions for disaster recovery (in case something happens on-premise) and availability to each service/location


Data Analytics Maturity
Company: Dental Magic; Sector: Dental Products
Current
Desired
Big Data Strategy
3.00
2.50
2.00
1.50
1.00
0.50
0.00
Organisation Maturity
Analytics Architecture
Big Data Usage
IT Maturity

Analysis Plan, Project Team and approach

The approach to be taken to complete analytic analysis of the data presented

We would use the CRISP-DM methodology, one of the most used methodologies for data analytics project. The reason for this results from the site assessment. As the companies services have a current data strategy in silos, we need to understand the data the company host and needs. Since this methodology is heavily focused on understanding the data to answer a business question, it seem a logical choice.

Since HR is a support function, we would follow the following approach:

- We would start in a discovery phase. This phase would deal with planning the different workshop among silos and stakeholder to drive the business area focuses and innovation labs. This is a component of business understanding to get critical insights to know where to focus effort. We can, for example, discuss how we can identify termination or no termination for employees.
- We would then define hypothesis and assumptions to prove. This would give initial thoughts on analysis and give hints on which aspect to explore and which data could act as predictor to prove business cases. For example, we may define Manager, Pay inequities or Age as predictor of termination.
- The next phase is the data and scope. This is the definition of a plan for data acquisition to specify which data, from where and how long. For example, we can specify that we need Employee data on position, manager, salaries, location, gender, demographics, Dates of hire and termination, department, Recruitment Source, Performance and Engagement, Absence and Workload.
- Then comes the analysis planning where we assess the data quality, we link the business to data to get an understanding to finish with the data preparation. For example, we could start by exploring different classification models, define the best performance, apply the principal of parsimony to have a more simplistic model with the main predictors and evaluate the cost of such model.
- Key findings would need to be communicated. We would use the model to gather data points whether they prove the business case or not. For example, 20% of employees of Manager X will voluntarily terminate. Or if the average of Absence per years is higher than 5, contract will likely be terminated.
- Finally, we could assess the impact of deploying a model. For example, we could say, if we are matching pay inequities, we could reduce our turnover by 10%. If we are training Manager X, we could save X € in lay-offs and recruiting fees.

The timeline & tasks for the analysis to take place

We would need to find the right business case by finding the right question. We would need the support for the Chief People Officer (CPO) to act as our main sponsor to understand the human resource and Talent manager data. We would need to conduct workshops to find where the data is located, who is going to coordinate the project with which team, on which focus area.

The panel of attendee needs to be a mix people involved in the company. We would mainly involved stakeholders from business side (Sales managers, Production Technicians II, Senior Accountants, Production Managers, Director of Operations) and IT side (Senior Database Administrators, Senior Network engineer, Data Architect, Data Analyst, Senior BI developer). We would also aim for mix of high level and low level to have the most balanced domain knowledge

possible and avoid having biased data. We would need to conduct as many workshops as needed to understand data and business sides to make connections we can come up with a team and provide specific business cases.

Once the business focus areas and analytics team are provided, we need to run a lab to prove the business case and use the organisation data to answer the business question. This phase will need to provide insights such as the assessment of the data quality, collaboration trainings, descriptive and predictive models, define Key Performance Indicators and final recommendations though regular Agile communications.

With this strategy, we would be able to provide a timeline. The time for it would be more than 8 weeks. Since each department is working in silos, there is additional challenges to provide a HR dataset, due to the sensitivity of employee data and authorisation required can be numerous, as employee may not be inclined to share their data. With silo architecture, extra time will be needed as gathering the data and accessing it will need a strategy to be defined and the participation of stakeholders in each silo. The data should not be too complex to gather. The main challenge resides in the speed and authorisation. A reasonable timeframe would aim with 12 weeks due to the organisation maturity and collaboration of employees and teams with each other.

Keep in mind, 83% of Analytics project fails, where fail means that data does not answer the business cases. This strategy may likely therefore need several iterations (other loops of 6-8 weeks after the initial 12 weeks) until we could deploy a model. However, we would understand what data would be needed to prove business cases, gather more domain knowledge or rethink which business cases we would solve.

The team roles to be filled

As said above, the first person that need to support the team is the main sponsor. For HR, involving and managing the expectations of the Chief People Officer is the crucial to drive such initiative and orient the business understand and decision making.

We would then need a domain business analyst, with a strong understanding of the domain dynamics and segmentation, and interest in analytics. This person would know how HR is driving business to understand and communicate on what good and bad look like. Brian Champaigne, the BI director seems to be the appropriate person to fill the position. Debra Houlihan, the Director of Sales could collaborate in this domain to central how sales work, as sales managers are dispatched individually across states.

We would then need a few data analysts and scientists to translate the business data into mathematics. These people should be very good at modeling, data mining and comfortable performing those tasks at the size of HR, meaning with every individual employee data. The company does not seem to have any senior data analysts or data scientists. We could train a few data analyst for data mining and research but it would be very interesting to hire a Senior Data Scientist temporarily to fill the gap in terms of building data models.

The Database engineer would help include a newer data lake technology. Experience in Database Admin and systems are critical for this task to make the data available for the project. Thanksfully, the company can count on Claudia N Carr, the Senior Database Admin and Katie Roper, the Data Architect to pilot this matter.

The UI developer would translate the story of the analytics solution. Their task would be to create every visuals to understand what the data says. For HR matters, a data analyst would deal with this matter as their does not seem to be any need for know to develop any front end or mobile application to gather data. If there is a business case needing development, however, we would seek external help.

The Project Manager needs to be cheering for effort and have the communication skills to facilitate dialogue between the team, business users and the innovation board. The company

seems to have a decent number of managers between Sales, Production or IT management to dedicate one to this matter. Depending on the business case, it would be smart to choose according to skills or domain knowledge.

Finally, the team would need a compliance officer. This person would be responsible for the data to be handled according to ethics and regulations. This person needs to have a deep understanding of GDPR and CCPA as well as AI act. We would not want our project to fail because of compliance or regulation changes. There does not seem to be anyone dedicated to this matter within the company, so we may seek external help here. Especially, an expert with fresh eyes could be less biased and have a critical view of the company, driving improvement more effectively.

Data Analysis

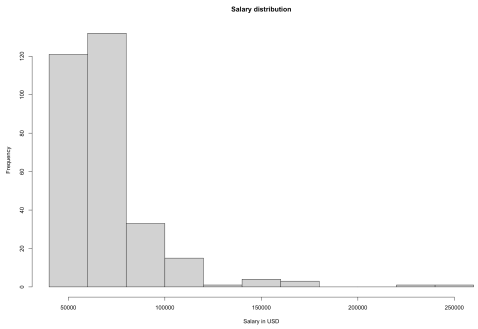
Data Quality Assessment

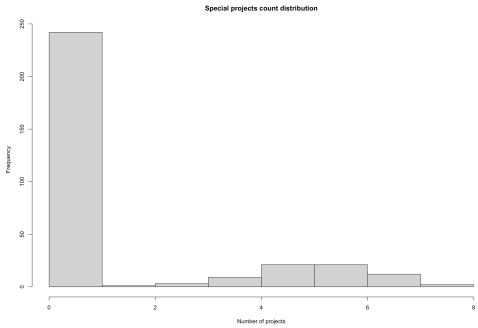
The data provided to us comes in the form of an Analytical Base Table (ABT) with 311 observations (rows for each employee) and 36 variable (columns). Data includes information about demographics (Name, IDs, Gender, Ethnicity, Date of birth, Location), Position (Salary, Department, Manager) and Performance (Score, Engagement, Lateness, Absences). We are dealing with mostly with factors and categorical variables (Names, positions, IDs, ...). The only numerical variables are Salary, Engagement Survey score (although typed as string of character), Employee satisfaction, Counts of special projects, lateness and absence.

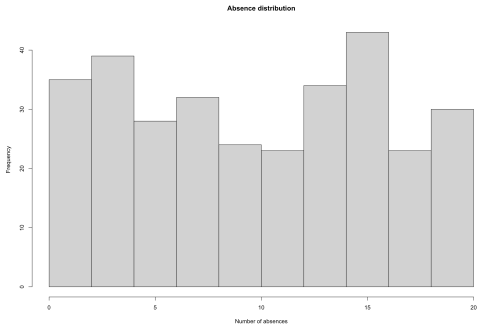
We can see the distribution of Data is right skewed, plotting an histogram, with a median salary of 62,810 USD and a maximum salary of 250,000 USD. Regarding the special project distribution, we have an isolated peak of almost no project, but we can notice a very slight peak between 4 and 6 projects. Absences seems bimodal with peaks at 3 and 15. We would note that absence counts are not dependent of years of employment, we would need to analyse this deeper. Finally, lateness records mostly counts of No Absence. Removing the 0 count makes a flat histogram.

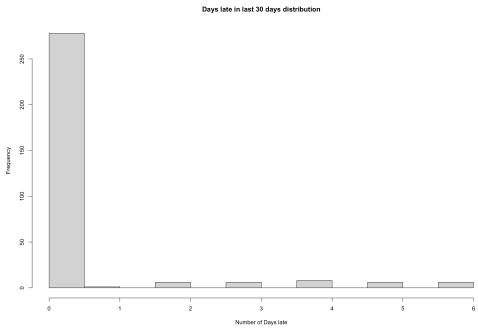
We are discerning a lot of outliers in Salaries. Using a boxplot, there is a significant number of outliers perceiving more than 100,000 USD, There is an obvious salary discrepancy but we will need to study those outlier as it could be a separation between managers and technicians.

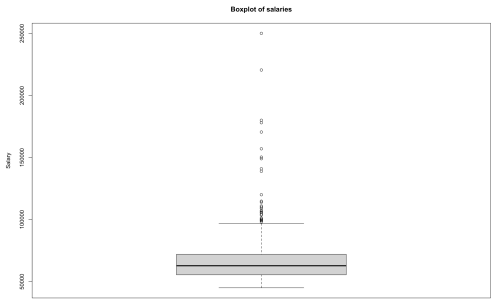
Data is quite complete. 207 missing values in Termination date but it makes sense as active employees are not terminated. 8 missing values in ManagerID but those can be mapped using Manager column. We are offered full values with factors for variables of interest (ie: EmpID correlate with Employment status, similar to Demographics).











Data Overview (MECE)

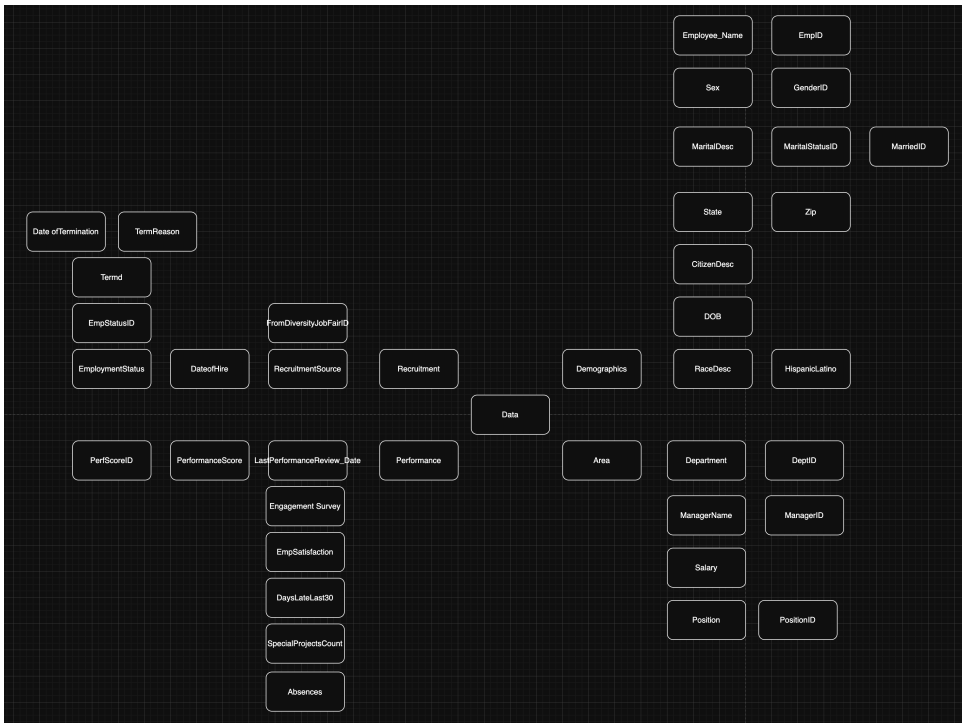
The data provided can mostly be separated into mutually exclusive and collectively exhaustive subset. We can distinguish 4 main subsets to separate the data: Recruitment, Demographics, Performance and Area of work.

Recruitment:
The recruitment contains information about the source and give an ID for Diversity Job Fair specifically, where 1 means employee comes from such fair and 0 from another source (Linkedin, Indeed, ...). It also contains the date of when the employee was hired and its employment status. Since this status contains different state of termination (cause, voluntary), an ID is provided where 1 refers to termination and 0 for still active. From then, we get the date of termination and container for reason. We can derive calculated field to get the employment period. Data seems self-explanatory and consistent.

Demographics:
This subset contains data about the employee as a person. We have the name and ID to give a unique identifier in the dataset. Information about sex and gender seem redundant at first (1 for Male, 0 for Females), but ID may be used for eventual models. Information about marital situation is giving ID for either a situation itself (Married, Divorced, Separated, Single, Windowed) and 1 for being currently married and living under same roof (1 being married). This helps separation if we want to study the non-married as group. We then information about location (State and Zip Code), Citizenship description, Date of Birth and Ethnicity with a factor on Hispanic (1 being Hispanic and 0 for not being). It would facilitate the difference, similar to marriage above. It is difficult to say if the data 2 or more races is MECE or not (we would need to know if parameter was strict when recorded)

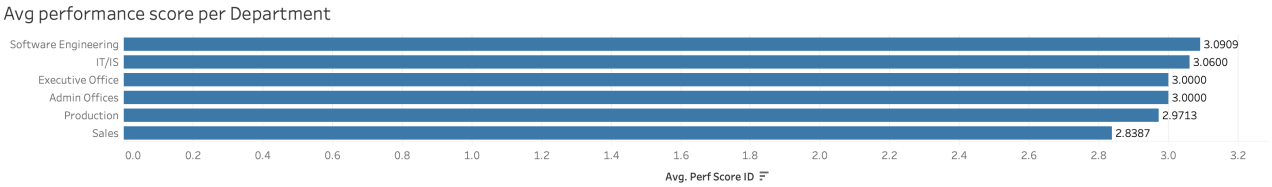
Area:
This information stores information the employee department, manager, Salary and position. ID seems a bit redundant but could be used for modeling similar to above.

Performance:
The performance variables are numerous. Except a performance review date that can be associated with a performance score and score ID, other variables cannot for sure be associated with each other. We have data about Lateness, Project Counts and Absences. We would also need to dig deeper on what are the difference between the score of engagement survey and employee satisfaction. We will assume that the employee satisfaction is a rating provided by manager and the engagement from the employee. Term reason is not MECE as we can see some entries are not exclusive (For example: the reason can be getting another position, but because employee was unhappy and/or wanted more money).

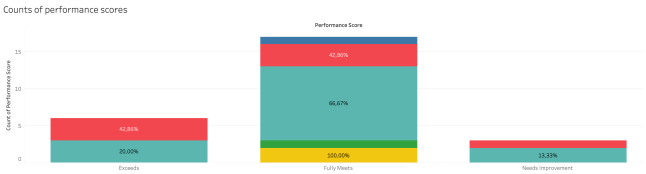


Is there any relationship between who a person works for and their performance score?

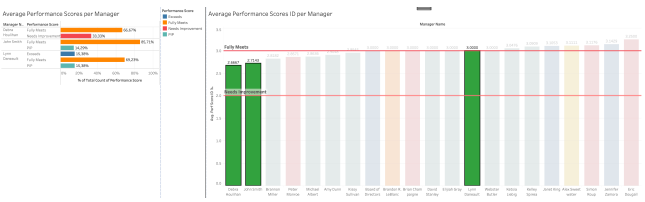
One business question is to understand whether or not a relationship exists between a manager and an employee performance score, the manager input affects significantly the performance score rather than other variables like department. (Appendix 1)



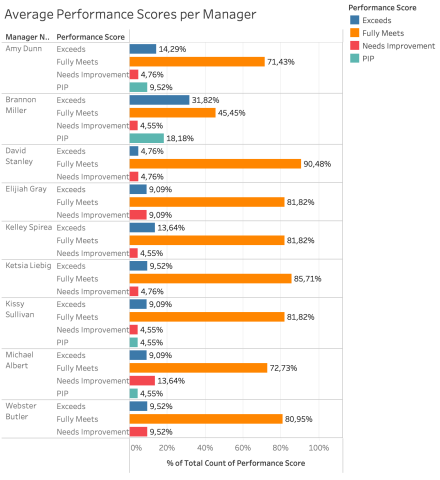
If compare averages by department, we can a hierarchy but all are very close to 3 (fully meets). Software engineering and ITs are a little over 3, Admin and Executive offices have an average of exactly 3, Sales and Production are slightly below. There does not seem to be a big difference among departments.



Let's address the managers that acts in different departments. Janet King, the CEO is managing in every department except in the admin offices. Based on data, we can see that she can provide needs, meets and exceeds in performance scores for production. She gave meets for most departments but also an exceed to 1 specific IT. Jennifer Zamora manages a hybrid team of ITs with 1 software engineer and we can see that she gave exceeds, meets and needs looking normal. We would need to look a little closer for the reason but we may argue that both managers could be comparing employees between themselves to provide a score.



We comparing average score ID per manager, we can see that there are difference in some manager within the same departments. If we take sales as an example, we can see that Lynn Daneault has an average of 3.0 (3 being fully meet) in comparison to the 2 other sales manager that have an average of 2.67 and 2.7143. Debra Houlihan manages a smaller team (3 employees) in comparison to other team of 13-14 employees, meaning any changes distorts the mean, therefore not the best point for comparison. John Smith has a similar team than Lynn and we can see John did not rate any exceed but give the same number of PIP (the lowest score). John may be less likely to give exceeds to employees.



Production, the biggest department in terms of employee number (209) has 9 dedicated managers for team of approximately 20 and we can observe slight differences in there averages ranging from 2.81 to 3.09. We can straightly see that Brannon Miller gave much less Meets rates (45.45%) while other team have more than 71%. He is also the manager that gave the most Exceeds and PIP ratings. In terms of Needs, Michael Albert have the highest average with 13.64%.

Based this data, there is evidence to argue that there is a

manager effect on the performance score. We would need to get an understanding on how the manager rates the employee, how the performance is assessed and whether or not each employee performance is compared with another to determine final rating. It would help to make the decision if manager or employee need trainings or if the performance scoring process needs to be rethought. We also need to determine if behaviour evolved on time.

What is the overall diversity profile of the organization?

Based on the data of active employees (Appendix 2), it appears there is a mix of male (43.96%) and female (56.04%) employees. The organization includes employees from various racial backgrounds, including White (59.90%), Black or African American (24.64%), Asian (9.66%), Two or more races (3.86%), and American Indian or Alaska Native(1.45%). The most prevalent group appears to be White, but there is representation from other groups.

A large number of employees are US Citizens (96.14%). There are also some Eligible Non-Citizens (3.38%).

A mix of single (48.79%), married (37.20%), divorced (6.76%) and widowed (1.93%) employees.

We can also calculate the age of employees and classify them in generations. At the company sizes, we can see that the company is mostly composed of Millenials (born 1981-1996, 51.22%) and Gen X (born 1965-1980, 43.96%) and some Boomers (born 1965-1979, 4.83%). We did not records any Gen Z in the data provided.

Keep in mind that we have missing variables to determine the diversity profile. There is no records on other factors like, disability, veteran status, religion or sexual orientation. Nowadays, we may also recognize non-binary genders and transgender so we need to keep in mind, the data does not allow to map a complete profile. A good or bad profile is difficult to determine so this question needs to be clarified according to the company culture.

We are attempting to compare by department, we can se a similar mix between genders, marital statuses and Ethnicity. We can see, despite Whites are the most represented in the company, there is no position or department that are fully represented by whites on big teams (more than 5 people). Similarly, marital status does not seem to be affecting the position. Generations seems to be mixed similarly to all departments.

Solely based on the data, we could say that organization has a somehow diverse profile, with age, marital status, ethnicity and gender bing spread across position and department. The decision whether it good or bad needs clarification. The company could determine their approach by conducting initiative to promote diversity and inclusion in the hiring process but in promotion and employee development. Seeking employee input and satisfaction could help determine if the question needs to be a priority to the business, as a lack of diversity could result in blockers or lack of perspective in terms of business strategy.

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

The data (Appendix 3) on activated employees shows that Indeed and LinkedIn are the 2 sources of recruitment among all employees, where 28% of employees are coming from Indeed and 24% from LinkedIn. Google Search comes third with 15%, leaving the rest of the share mostly among CareerBuilder, Job Fairs, Employee referrals and company website.

An interesting fact is that only Indeed has a take on every department, meaning it is the only source that records at least one instance of each department. Another interesting fact, is that the Diversity Job Fair only provided Black and African American people only. Although it can be a reliable source of diversity for the company, the area is worth reflecting to understand such an absolute number. There may be a bias in the recruitment or in the audience attending job fairs. The job fairs provided almost every department, with the biggest rate in Admin (33%) and Software Engineering (27%).

Regarding Gender, we can see that both of the biggest providers (Indeed and LinkedIn) are sourcing 42% of Males and 58% of Females for the company. Google Search seems to source a bigger rate for Females (67%). The only source sourcing more Male than Female Employees are the employee referrals.

Finally, looking at the distribution between departments, it seems at first that the distribution seems similar across the biggest sources. Each is allow most of their share to production, meaning production technician are likely to be hired from those 6 sources. Sales agent are mostly coming from Indeed and Website. There is a significant shares from ITs coming from Indeed but also from Employee referrals and LinkedIn.

Similar to previous question, the diversity aspect needs to be defined to determine what is the best source. Diversity can be at a site, department, or position level, so it is very important to define the culture around this aspect before actually making a decision. Due its wide range in terms of Ethnicity and application to department, we could say Indeed seem the best source to cultivate diversity. Diversity fairs may be used if the company needs more people with a Black or African American background, and LinkedIn seems to also provide a fair share of backgrounds. If more Male employees are needed, a campaign of Employee referral could be a good solution to increase numbers as it is the only source providing more Male Employees.

But again, before making actions, clear goals around diversity need to be defined. Demographic data gathered seem very limited in comparison to the diversity standards of an American company expected nowadays.

Can we predict who is going to terminate and who isn't? What level of accuracy can we achieve on this?

The company hired its first employee in 2006 and the first to terminate in 2010. (Appendix 4) Deriving a calculated field using dates of hire and termination from data, active employees have been at the company 11.76 years on average with a standard deviation of 1.94 and employees spend on average 8.95 years at the company (standard deviation: 4.42). The records indicates 2 states for termination either for cause (company laying off the employee) or voluntarily (employee resigned). On average, terminated employees have spent 3.36 years (standard deviation: 1.99), where termination for cause happens on average after 3.18 years (Standard deviation: 2.04) and voluntary termination occurs after 3.39 years (Standard deviation: 1.99). The first 5 years are therefore important to monitor to reduce termination, looking at those numbers, We may argue the less likely to terminate are employees between 5 and 9 years of employment.

Looking at counts, we can see 207 active employees, for 16 terminated for cause and 88 voluntary termination over the years. Among causes, we can see most are related to attendance (6 for absences and 4 no-call, no-show) and performance (3). For voluntary termination, it a little more difficult as the data is not MECE. Main reasons include getting another position (20), being unhappy (14) and needing more money. As you can see, these reason can be either correlated or causing each other (employee got another position because unhappy and asking more money). Career change, working hours are other important aspects but it is difficult to predict who is going to terminate based on this.

We can attempt to perform a logistic regression and experimenting with predictors to classify TermID (0 being Active, 1 being Terminated). After experimenting with this data, we can derive a logistic regression model using Absences, Lateness and Special Project as predictors. With a 70% training-30%testing split, we can have model with an accuracy of 30%. This accuracy score is naturally not an acceptable threshold to use the model. It is may be interesting to gather more data and make it MECE to strengthen our predictors.

If counting average lateness by department, Production records the highest average for each employment status category. Digging more into position, the highest averages goes to Production Technician I. Finally, we can see that the distributions of each employment status differ. We can see that Active employees have their year of employment normally distributed but containing outliers on both sides of the plot (meaning too fresh or too tenure in comparison to the other employees) while Cause or Voluntary termination are skewed.

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

Using Boxplots (Appendix 5), we can see inequities by departments. We can see explained inequities in Admin offices with a right skewed distribution of the salary, explained by the level of seniority (Senior accountant earning more than accountants, earning more than assistants). Sales follows a similar distribution based on area width, only the Head of sales is an outlier. Software engineer do not display any outlier and seem normally distributed with salaries ranging from 77,692 to 108,897 USD. We do not have the data on skills but we may assume that salary may be based on skillset.

The production seems normally distributed but displays a few outliers and an extreme outlier at 170,500 USD, very far from IQR (53,018 to 64,066). These discrepancies are explained with the level of responsibility. The extreme outlier is the director (head operation) and the 3 outliers (closer to crowd than to director) are Production managers.

It gets interesting with IT. While salaries are normally distributed, we can see a very wide IQR (76,029 to 107,226) but also wide whiskers (50,128 and 150,290). Some ITs are perceiving 3 times the salary of least paid one. We also have 3 outlier, where 2 are explained with the level of responsibility (CIO and Director of IT), while the other is the Infra manager. We would need to dig deeper in to positions. We can see that IT involve different position based on skillset. For each position, we are not seeing any outlier, except one Sr. Network engineer paid much more than the other Srs (107,226).

When viewing salaries against gender, we can see than median salaries for Male and Female workers are not far from each other, but Female distribution seem normal while Male Distribution is right skewed. The upper whisker for Male, however is significantly higher than Female's (18,223 USD of difference). We can also see outliers in both genders. Male outliers are explained by skills or responsibilities (ITs, Directors, Managers) similarly to Female. Extreme outliers are explained with their position (CEO, CIO, Head of sales.

Finally, Ethnicity displays some pay inequities. We can see that White employee have a narrow IQR of salaries in comparison to Asians or Black/African American. Inequities seem all explained the same way, productions have the lowest salaries, then comes sales and higher salaries to managers or ITs. Asians do not display outlier but Black/African American contain outlier paid higher, which is explained by their position in the IT department. Finally, White employees contains a significant number outlier, explained by their seniority, belonging to IT or Directors. We can see that the directors are split between White and Black background.

Conclusions

Dental Magic has bought the idea Big Data can make an impact on piloting the business. It seem that the company has the capacity to launch analytics in terms of skillset but needs a strategy. Most areas they need to work on deal with the realignment of the communications as business seem to work in silos, involve Data scientist in their workforce but also implement compliance with Data regulation.

Due to its strategy challenges, especially the silos, we recommend a timeline of 12 weeks with a dedicated team for a CRISP-DM methodology. As the company is new to Big Data and speed of getting authorisation and access to data may be time-consuming, it will take time to iterate between the data understanding and defining business cases.

The data comes in the form of ABT and contains data on demographics, performance, recruitment and position for each of the 311 employees. We can see, with some examples that some distributions are not normal but data seems to come in complete. Missing values are either few or explained.

The data provided is a good start to investigate the human resource side of the company. We can see very few missing data and logical split between demographics, Recruitment, Performance and Position. However the data quality is not perfect, since we can find that some aspect are not MECE (Termination reason).

Data Findings and Next Steps

After analysis, we may argue a relationship between the performance ratings and the manager. Performance scores do not appear to be depending on department or position variable and we can see the average between managers are not so different. However, we noticed that some managers tends to give more exceeds or PIP than others. We took Production managers apart as their teams have more employees and we can see than some manager are more prone to give exceeds and PIP in comparison to Meets, some are only giving meets and others are seem to distribute normally (a few needs and exceeds, mostly meets). Depending if the company is seeking alignment, the next step would be investigating the reason of ratings (are those based on objectives or are employees compared between each other).

The data provided does not allow to give a full diversity profile. We are missing data about religion, disability, military at least and nowadays non-binary gender or not disclosing are values that could be mapped for gender. Moreover, we do not have an idea of what is a good or bad diversity profile. Based on the data gathered, it seems the company has diverse profile for an American company with a mix of gender, marital status, age and background. The next step to address the business question would involve gather more specific data and give a better definition in terms of number for a diversity profile. The company could investigate other companies or pilot initiatives to seek employee input. It could be seen as a satisfaction driver.

In the same line, as diversity is not clearly defined, we can only make assumptions about what is the best recruiting source to cultivate diversity. We saw that Indeed seem to have sourced at least one person for every department. We also pointed that diversity fair sourced only Black/African American people. Employee referrals mostly source Male employees, while other sources usually provide female employees 2 out of 3 times. Depending on the diversity objectives, the next step to progress towards this business question would deal with the definition of the profile. Other areas to investigate are why the diversity job fairs only provide Black backgrounds but also think about a strategy that would come around Indeed as main source, since it provided workers for every department. Linkedin seems reliable and the company could launch an employee referral campaign in case more Male employees are needed.

It is difficult to predict who is going to terminate and who is not. The data does not allow to investigate the reason as this variable is not following MECE principles. We can only say that Absence, Special Projects Counts and Lateness are good predictors to predict but building a model with the current data would result with a very poor accuracy (30% not acceptable to deploy). Using averages, we would need to watch the absences and lateness within the first 5 years of employment as the average years of employments for terminated contracts is 3.36 years. The next step here would deal with investigating reasons more thoroughly to make the variable MECE and investigate reason for lateness or absences. We may be able to use a reason as predictor for a model such as Commute, Sickness, Motivation or other factors.

Finally, we distinguished a few pay inequities within the company. Most inequities are explained by the level of seniority, the department, the skills set or level of responsibility. However, we could pinpoint some outliers in IT due to its position and skillset but some ITs with similar positions have significant salary differences (up to 3 times the salary of the least paid IT). If we make a gender comparison, it seems that Men and Women do not follow the same distribution as Women salaries are normally distributed while Men salaries are right skewed. It can be argued a few instances of Men are better paid than Women in similar position. The salary gap should be looked at to resolve those inequities. Finally, Ethnicity is to be considered as Black employees have a wide IQR and a significant number of outliers among themselves while White employees have a narrow IQR. It is also worth noting that top 3 salaries are owned by Whites. The next steps would be to resolve the discrepancies between Men and Women and among Black employees. Since IT has a wide range of position, it would be also worth digging the compensation away from skills as it generate discrepancies between each area of expertise.

In conclusion, the data as is gave us a solid understanding of HR and allowed use to make some actionable decisions such as investigating and streamlining the performance scoring process or point a few areas to resolve pay inequities. However, we need a better understanding of the data

and the company culture to define a diversity profile or predict termination. We need an alignment and more quality and granular data for those purposes.

Appendix 1 - Performance Scores

Fig 1 - Performance scores distributions

Appendix 2 - Diversity Profiles

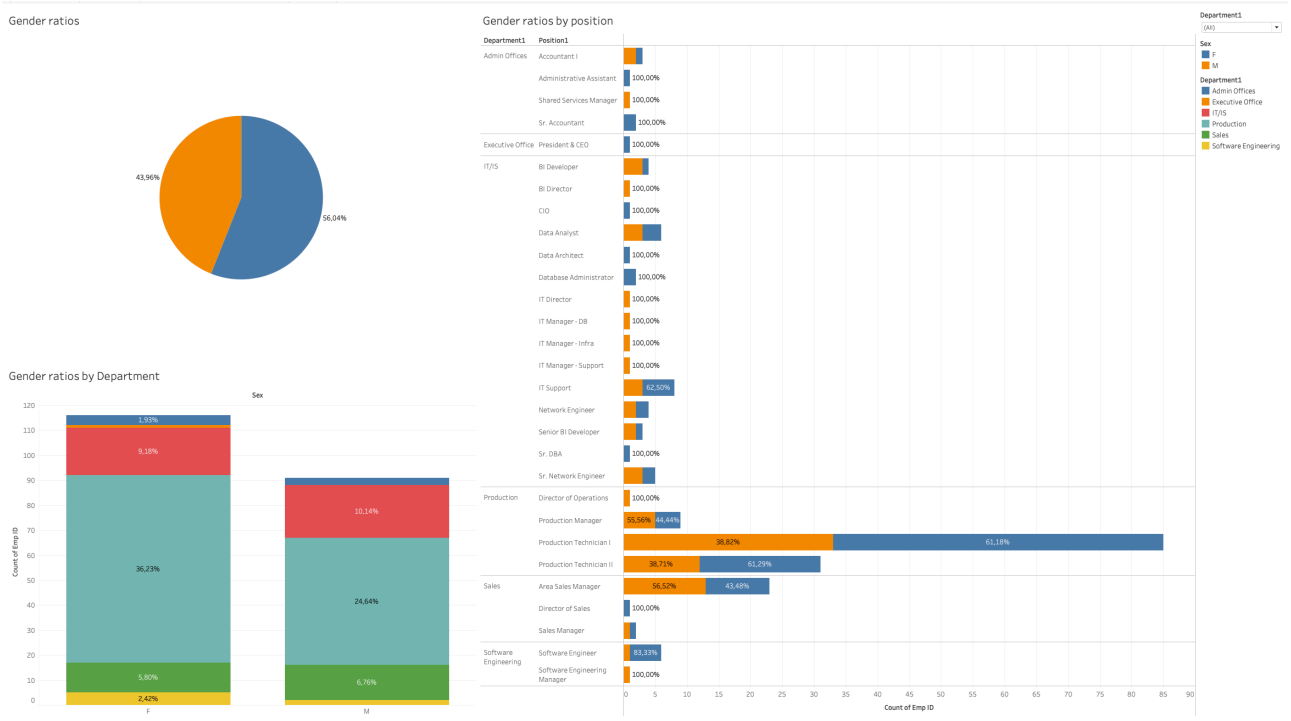


Fig 2 - Gender distribution

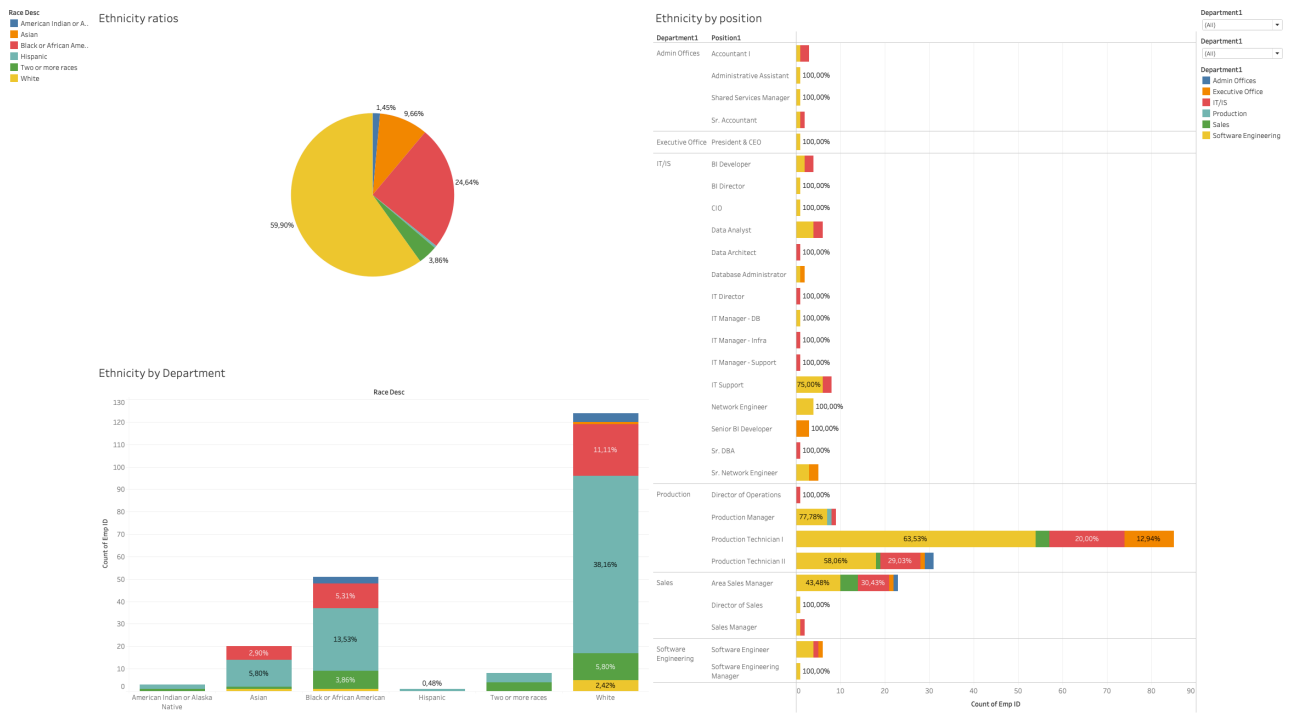


Fig 3 - Ethnicity distribution

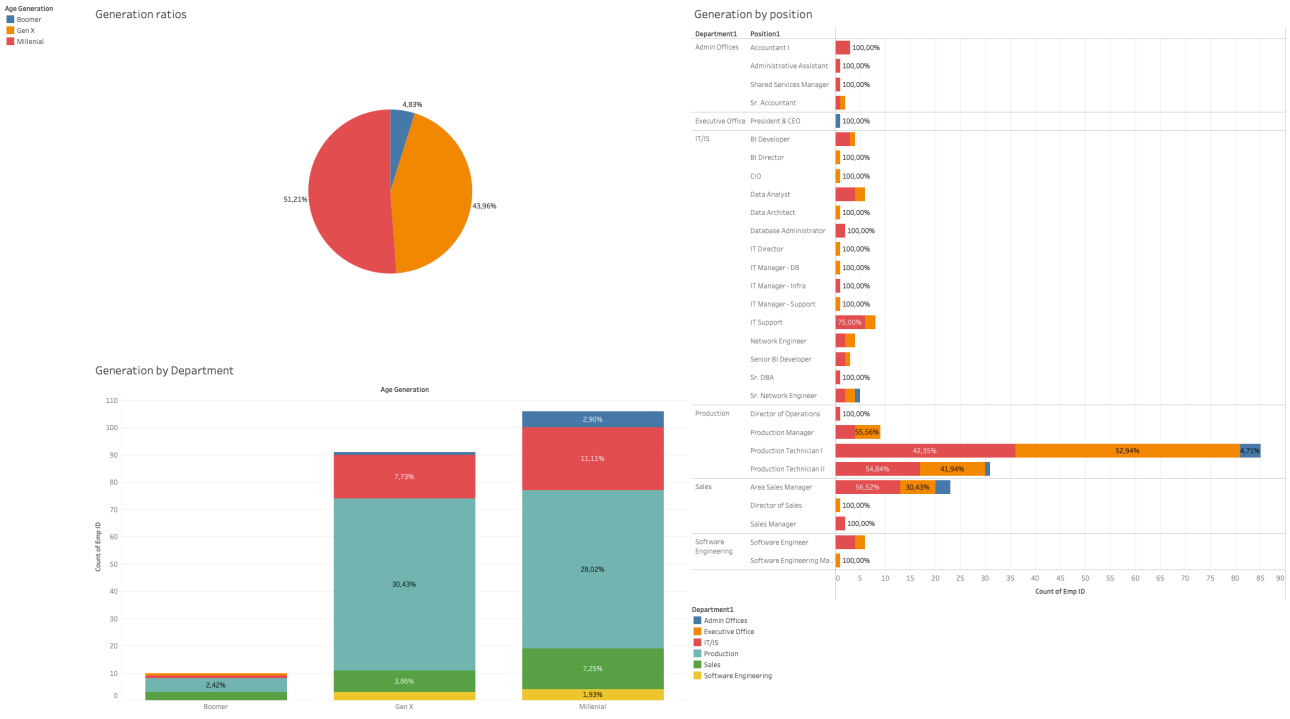


Fig 4 - Age Generation distribution

Appendix 3 - Recruiting Sources

Fig 5 - Recruiting Sources distribution

Appendix 4 - Terminations

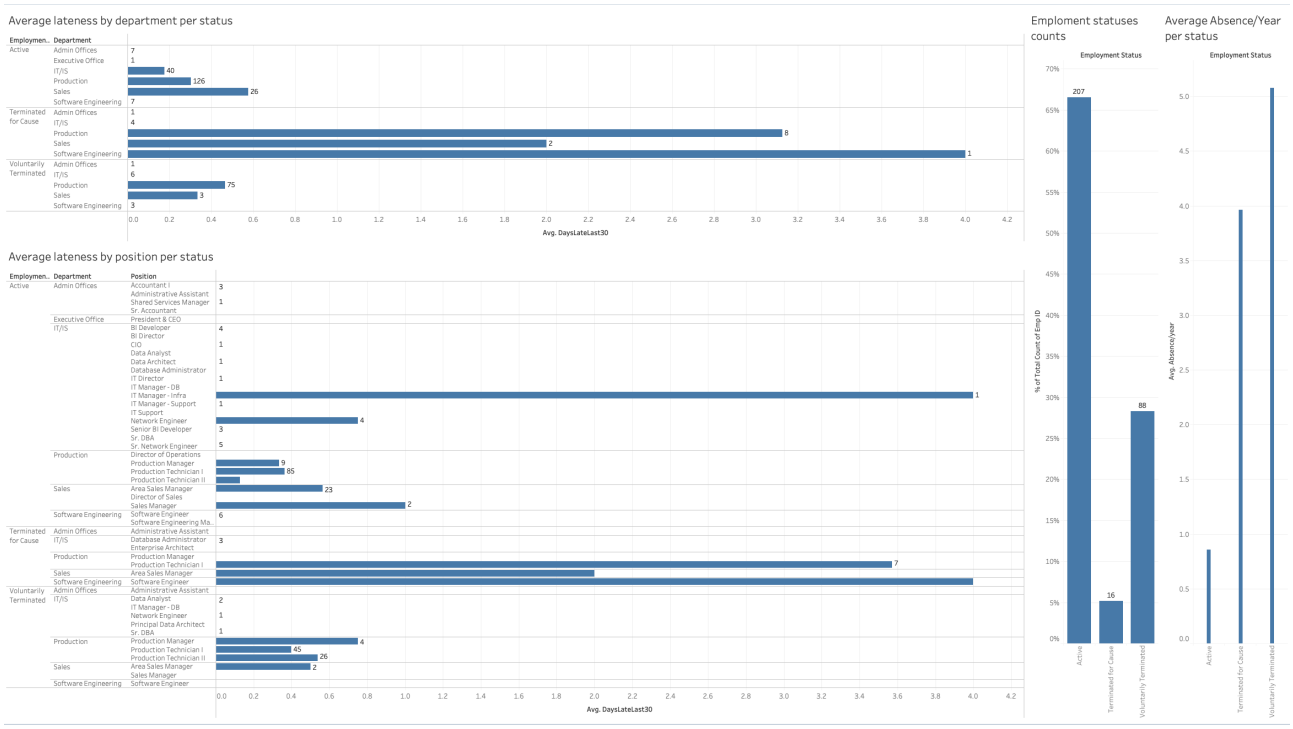


Fig 6 - Recruiting Sources distribution

Appendix 5 - Pay Inequities

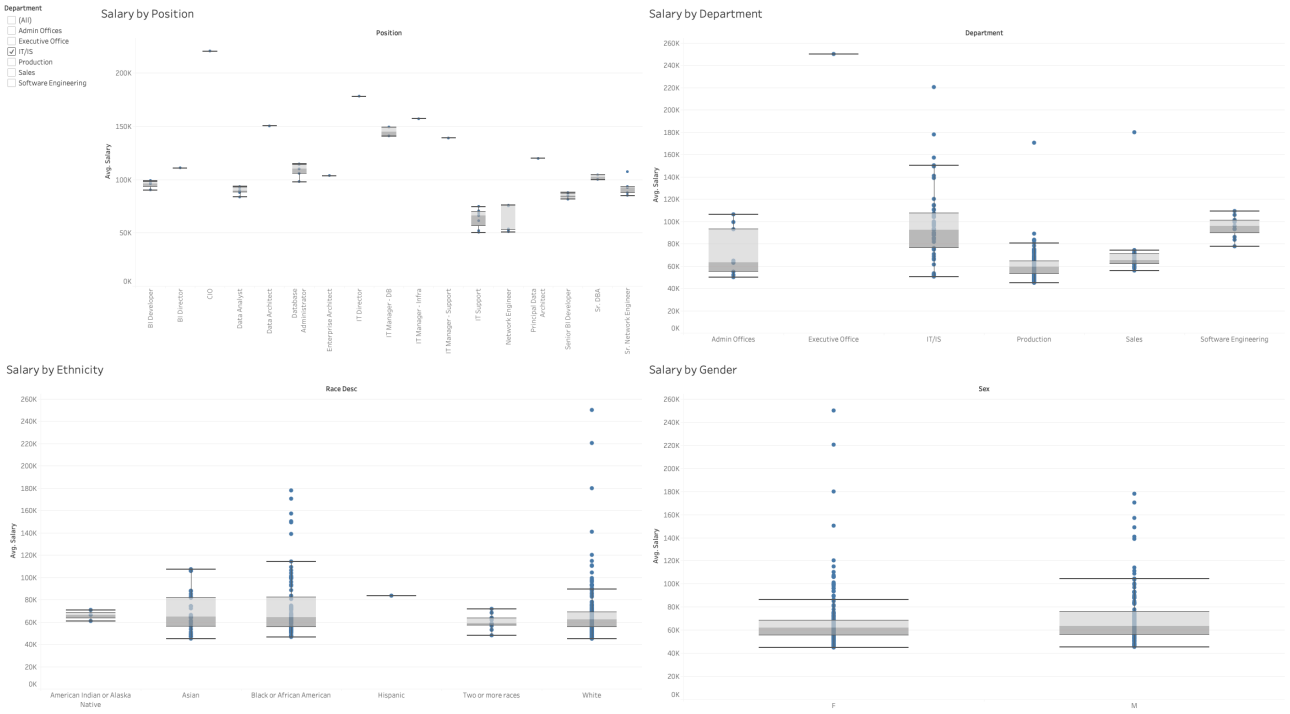


Fig 7 - Salaries distribution