UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Trabajo Académico Visualización

"Dashboard Para Recursos Humanos"

Realizado por Andrea García Pastor y Elizaveta Gilyarovskaya

Grado en Ciencia de Datos

Fecha 27/01/2020

Índice de Contenidos

1. La pregunta

Staff attrition es un dataset ficticio creado por IBM. Attrition o la pérdida del personal se refiere a la pérdida de empleados a través de un proceso natural, como jubilación, renuncia, eliminación de un puesto, salud personal u otras razones similares. El término inglés attrition se refiere al caso en el que un empleado se retira y la empresa no planea cubrir la vacante dejada por el ex empleado, aún sabiendo que va a producir un efecto negativo..

Dada una empresa ficticia, nos planteamos la siguiente pregunta: ¿qué recomendaciones se le podría dar para reducir la tasa de pérdida de personal y así, evitar el efecto negativo que produce esta? Se aboradará, en primer lugar, mostrando las características generales del personal y posteriormente, las características de solo aquellos empleados cuyo valor en la variable Left sea afirmativo, es decir, los que se han ido.

2. Fase de Adquisición

La etapa de adquisición de datos ha sido bastante sencilla una vez elegido el dataset sobre el que se ha decidido trabajar, que ha sido la decisión más complicada de todas con diferencia. Se trata de un csv descargado de Kaggle con 35 variables (la mayoría categóricas o numéricas discretas) y 1470 observaciones.

Referencias: [https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset]

3. Fase de Formateado

Después de adquirir los datos y hacer el análisis exploratorio, es necesario modificar el formato de algunas variables del dataset. Las transformaciones generales realizadas han sido:

Renombrado de variables:

```
employee = read.csv('WA_Fn-UseC_-HR-Employee-Attrition.csv')

names(employee)[2] = "Left"

employee$Department = mapvalues(employee$Department, from =
                      c("Sales","Research & Development",
                      "Human Resources"), to = c("Sales","Res$Dev","HR"))
```

Creación de una nueva variable a partir de Age:

```
employee$Generation <- ifelse(employee$Age<30,"x<30",
                       ifelse(employee$Age>=31 & employee$Age<55,"30<x<55",
                       ifelse(employee$Age>=56 &
                              employee$Age<70,"55<x<70","x>70")))
```

Ordenado de los factores dentro de esa nueva variable:

```
employee$Generation= factor(employee$Generation,
                            ordered = TRUE,
                            levels = c("x<30","30<x<55","55<x<70","x>70"),
                            labels = c("x<30","30<x<55","55<x<70","x>70"))
```

4. Fase de Filtrado

A pesar de haber muchas variables, no se ha eliminado ninguna ya que no se sabía desde el primer momento cuáles acabarían teniendo más o menos importancia para responder a la pregunta. Algunas de las que se ha prescindido son: EmployeeCount, Over18, BusinessTravel, DailyRate, DistanceFromHome, Education, EducationField, HourlyRate, JobInvolvement, JobLevel, MonthlyRate, NumCompaniesWorked, OverTime, PercentSalaryHike, PerformanceRating, StandardHours, StockOptionLevel, etc….

Para la segunda pestaña del dashboard se han eliminado todas aquellas observaciones cuyo valor en la variable Left (la que representa attrition) sea negativo con el fin de ver las estadísticas de sólo aquellos empleados que se han ido.

5. Fase de Minado

La parte de minado de datos se realiza en las primeras capas de los gráficos plotly, de forma personalizada para cada gráfico ya sea sacando grupos de datos o estableciendo límites de dibujado. Por norma general lo que más se ha empleado ha sido el group_by para sacar un conjunto de datos necesario, guardarlo en una tabla y aplicar summarise para sacar estadísticas dentro de este conjunto, etc. Por ejemplo:

```
env_attr = employee %>% select(EnvironmentSatisfaction, JobRole, Left) %>%
          group_by(JobRole, Left) %>%
          summarize(media_amb=mean(EnvironmentSatisfaction))


p <- ggplot(env_attr, aes(x=JobRole, y=media_amb)) +
      geom_line(aes(group=Left), color="#8946A6", linetype="dashed") +
      geom_point(aes(color=Left), size=3) + theme_tufte() +
      theme(plot.title=element_text(hjust=0.5),
```

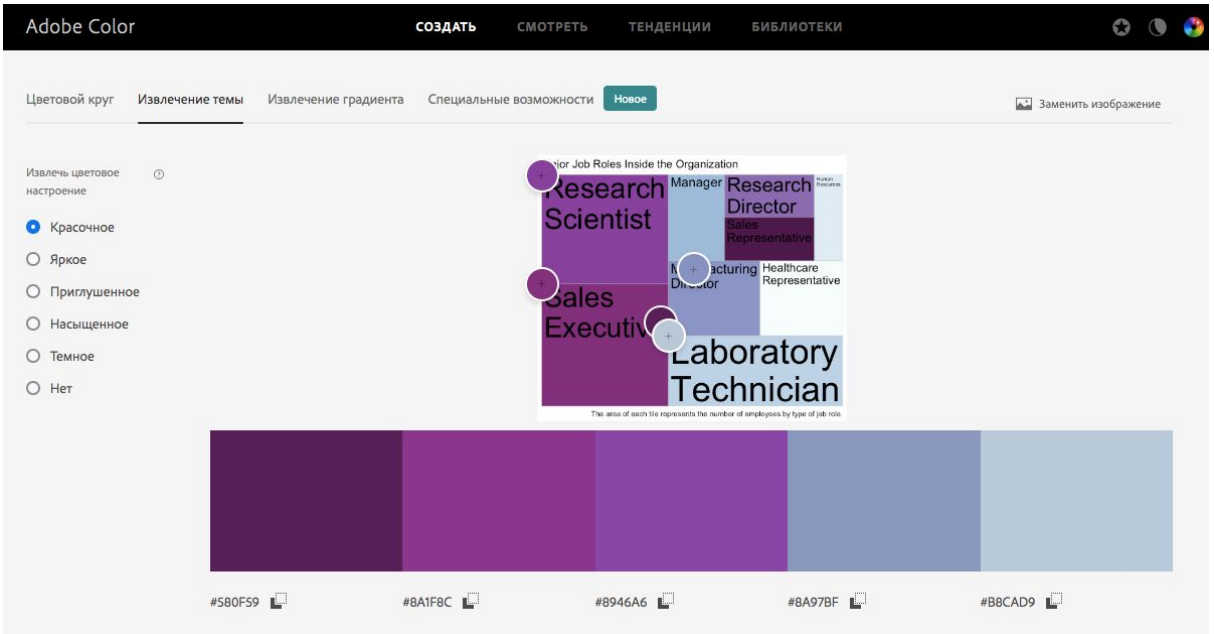
```
axis.text.x=element_text(angle=90)) +
      labs(y="Satisfacción de Ambiente Media", x="") +
      scale_color_manual(values=c("#580F59", "#8A97BF"))
```

6. Fase de Representación

A la hora de representar los datos, se han utilizado diversos tipos de gráficos, pero todos bidimensionales. Sin embargo se ha intentado evitar los gráficos de tarta por su ineficacia y se ha optado por los gráficos de barras en la mayoría de las ocasiones dada la naturaleza de las variables (casi todas cualitativas o discretas con pocos valores). El color se ha utilizado como el canal para establecer diferencias entre los factores de las variables categóricas.

7. Fase de Refinado

En cuanto al refinado, se ha utilizado una misma temática para todos los gráficos (theme_tufte) y utilizado una gama de colores monocromática de morados y azules ("#8A208C", "#8946A6","#8A97BF","#A0BED9"). No se han empleado efectos visuales tales como sombra o 3D.



Gracias a esta paleta se ha podido añadir un poco más de detalle en los gráficos, representando los datos según el nivel de satisfacción y el género, entre otros aspectos.

Además, todos los ejes X de los gráficos se encuentran ordenados numéricamente y alfabéticamente si los datos son cuantitativos o cualitativos, respectivamente.

8. Fase de Interacción

Para añadir una capa de interacción a los gráficos se ha utilizado tanto plot_ly como ggplot (pasándole la función ggplotly).

Si se mueve el puntero a una columna, a un pico o a un punto, se muestra el valor que representan e incluso información adicional si así se ha programado, ya sea el nivel de satisfacción, el género o incluso la mediana, los cuartiles, el máximo y el mínimo en el caso de un boxplot.

También se permite centrar los datos, es decir, con un doble click sobre un elemento de la leyenda, el gráfico cambia solo para mostrar los datos pertenecientes a la opción u opciones elegidas.

Por otro lado, en el gráfico que representa el estudio por Generaciones y géneros, se ha añadido animación con respecto a los años que se ha trabajado en la empresa. Para ello, se ha hecho uso de highlight_key sobre la base de datos y de un filter_slider, un evento de filtrado que permite seleccionar un rango de valores numéricos mediante una barra de deslizamiento.