Prueba de Conocimientos – Científico de datos

1. A) Tengo una variable aleatoria x que no conozco su media pero si su desviación . Pero x representa la sumatoria de varios pesos y la variable aleatoria individual para cada peso se representa como x/6 por lo tanto la media es y la desviación de esta variable aleatoria es .

Por lo tanto, la media es:

B) Codigo en R

qnorm(0.021,lower.tail=FALSE)

x = 500/6

(desv = 15.7/36)

((qnorm(1-0.021)\*desv)+x) #84.22017

pnorm(84.22017,x,desv,lower.tail=FALSE) #0.021

S50<-rnorm(50,x,desv)

sum(S50>84.22017)/50 #0.06

S100<-rnorm(100,x,desv)

sum(S100>84.22017)/100 #0.02

S1000<-rnorm(1000,x,desv)

sum(S1000>84.22017)/1000 #0.019

1. Pienso que la tecnología más útil para esta información estructurada podría ser Amazon RDS (Relatinal Database Service) aunque podría ser interesante Amazon Neptune puesto que es para bases de datos Graph, que permite trabajar con datos altamente relacionados y es muy potente para resolver queries que en datos estructurados normales se vuelven muy complicadas con muchos join.
2. Pienso que la tecnología más útil para bases de datos columnares que permita ser eficiente en consulta y permita conectar sus consultas a un tablero de visualización es Amazon RedShift, puesto que esta tecnología esta basado en almacenamiento en columnas y esto es un factor importante en el desempeño de consultas analíticas, ya que reduce notablemente los requisitos de desempeño globales de E/S del disco, disminuyendo el volumen de datos que hay que cargar en él, reduciendo la escala utilizando clústeres distribuidos de hardware de bajo coste para aumentar el desempeño. También Amazon RedShift es rentable utilizando las herramientas de inteligencia empresarial existentes.
3. Tipos de visualización para cada variable:

episodioId: Podría aplicar o explorarse un grafico de barras de cuente los múltiples eventos de cada episodio, pero que permita visualizar los diez (10) primeros episodios con mayor cantidad de eventos.

eventoId: No aplica gráfico, puesto que es una variable identificadora de los eventos.

Departamento: Aplica un grafico de barras o barplot, donde se cuente los eventos.

fechaInicial: Individualmente este campo no aplicaría ningún gráfico, pero ya multidimensionalmente aplicarían gráficos donde se permita ver el comportamiento de otra variable a través del tiempo, por ejemplo, un grafico de muertes totales o lesiones totales a través del tiempo (time series plot).

torFscale: Un grafico de tortas, donde se represente el porcentaje de representación de cada escala de Fujita-Person, lo cual nos mostraría la intensidad más frecuente. Pero este grafico no permitiría ver la transición de la intensidad del evento, no se que tipo de grafico permita analizar esa transición.

ubicacionOrigen: Aplica un barplot que permita identificar dichas zonas con mayor cantidad inicios de los eventos.

ubicacionDestino: Aplica un grafico de barras que permita identificar las zonas con mayor finalización de eventos.

latitudInicial: Individualmente este campo no aplica ningún gráfico, pero en conjunto con la longitud inicial puede representarse en un gráfico de mapa.

latitudFinal: Individualmente este campo no aplica ningún gráfico, pero en conjunto con la longitud final puede representarse en un gráfico de mapa.

muertesTotales: Univariadamente este campo se podría utilizar los gráficos de histograma y boxplot. El histograma de esta variable permite ver una posible distribución univariada empírica de los datos.

lesionesTotales: Univariadamente este campo sirven tanto el histograma y el boxplot, pero es más útil el histograma para observar distribuciones univariadas.

impactoEstimado: Univariadamente esta variable se podría representar en histograma y en boxplot. Pero es más efectiva el grafico del histograma para representar la distribución univariada.

1. La respuesta correcta es la d, los clusters no proporcionarán información útil. Puesto que no se realizo un proceso de normalización, haciendo que las métricas más grandes tendrán mayor peso en la conformación de los clusters. En este caso el campo salario\_anual opacara al campo edad\_empleado, haciendo que el agrupamiento no sea muy útil.