Logotipo, nombre de la empresa

Descripción generada automáticamente

Universidad de Montevideo – Introducción a la Ciencia de Datos

Parcial 2021

13 de Octubre, 2021

**1.** Suponga que un Banco lo contrata como consultor para desarrollar un modelo que identifique qué potenciales clientes tienen alta probabilidad de no repagar un crédito (son más riesgosos).

1. Diagrame el workflow del proyecto con sus principales características.
2. ¿En qué etapa es más probable que invierta la mayor parte de su tiempo?
3. ¿Qué tipo de pregunta intenta contestar su cliente? ¿Qué otro tipo de pregunta le parece que agregaría valor al análisis y por qué?
4. ¿Recomendaría a su cliente utilizar datos externos? ¿Qué ventajas y desventajas tendría?

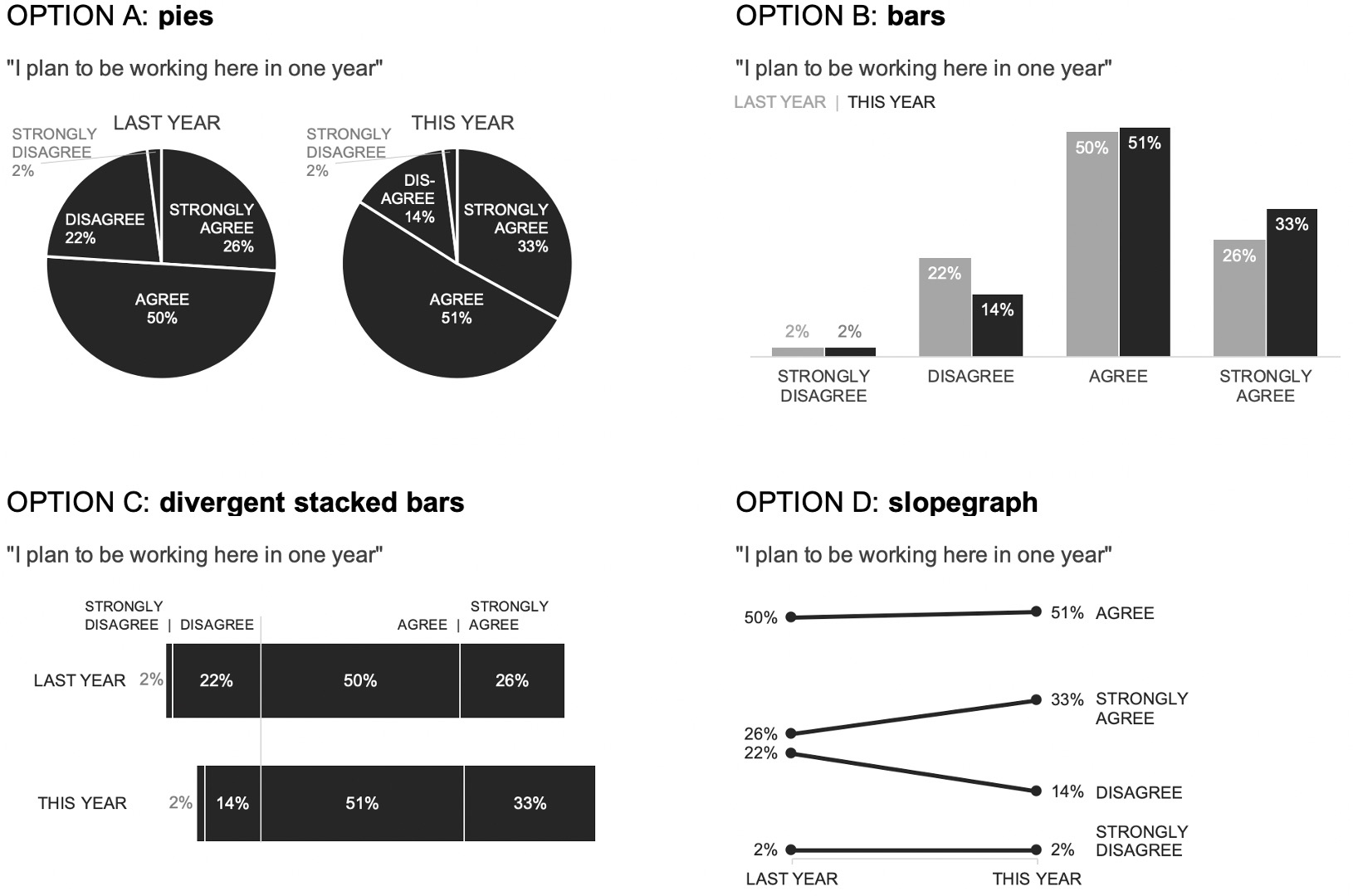
**2.** Estadísticas Descriptivas:

1. ¿Cómo influyen los outliers en las medidas de tendencia central?
2. ¿Qué medida de tendencia central se recomienda usar para datos ordinales? ¿Por qué?
3. Describa como espera que sea la relación entre la media, la mediana y la moda para las distribuciones: normal y right skewed.

**3.** Una máquina está configurada de manera que el contenido promedio por botella sea igual a 50. Una muestra de 36 botellas arroja un contenido medio de 48,5 ml. Suponga que la desviación estándar de la población es σ = 5ml.

1. Plantee y pruebe la hipótesis de que el contenido medio por botella es de 50ml al nivel de significancia del 5% (1.96).
2. Qué significaría en este ejemplo cometer un error Tipo I.
3. ¿Qué es el p-value?
4. En general, ¿por qué se utilizan muestras en vez de utilizar la población? ¿Qué precauciones debo tomar al hacerlo?

**4.** Imagínese que está visualizando los resultados que los empleados de su empresa contestaron a la siguiente afirmación: "Planeo trabajar aquí en un año" en una encuesta de satisfacción laboral. A continuación, verá cuatro visualizaciones diferentes de los mismos datos.



1. Enumere ventajas y desventajas para cada una de las cuatro opciones (deben aplicar al ejemplo concreto y no ser generales).
2. Suponga que tiene la tarea de comunicar estos datos. ¿Cuál escogería? ¿Le haría algún cambio? ¿Se le ocurre alguna opción diferente?

5. Usted ha entrenado un modelo no supervisado de clustering con el fin de segmentar su clientela en grupos. Como hemos visto en clase, la clusterización puede resultar en muchos casos ambigüa. Una vez fijado el número de clusters, cómo haría usted para validar que los grupos formados son correctos y tienen sentido?

6. Cómo y por qué se da cuenta a simple vista que la siguiente serie es no estacionaria? Qué haría para estacionar la serie?

Gráfico, Diagrama, Gráfico de líneas

Descripción generada automáticamente

7. Suponga que usted tiene un dataset con muchas columnas numéricas continuas en distintos rangos de valores y una columna categórica nominal con muchas categorías. Le solicitan que entrene un modelo no supervisado de clustering. Qué problema cree que puede surgir y qué preprocesamiento cree que debe hacer a las columnas para evitar dicho problema?

8. Suponga que usted es contratado por una compañía multinacional para realizar un análisis de sentimientos de su clientela utilizando los tweets relacionados a la compañía. Explique qué preprocesamientos debería hacer previo al análisis. Tenga en cuenta el contexto informal de las redes sociales.

9. Suponga que usted tiene tres tablas A, B y C. Entre A y C no comparten una key en común. Sin embargo, B es una tabla que captura la relación entre A y C por medio de una key compuesta por la key de A y la key de C. Explique como realizaría el JOIN de las tablas A y C.

10. Explique las cuatro características fundamentales de un buen científico de datos.