Logística de envíos: ¿Cuándo llega?

Mentoría DiploDatos 2019

Práctico: Análisis y visualización de datos

# Motivación

En la actualidad, cada vez más productos se comercializan a través de una plataforma online. Una de las principales ventajas de este sistema es que el usuario puede recibir el producto en su domicilio en una fecha determinada. Pero, ¿cómo sabemos qué día va a llegar? ¿A partir de qué datos podemos predecir la demora del envío? En este práctico se trabajará con datos de envíos de MercadoLibre, el e-commerce más grande de Latinoamérica, analizando y modelando el problema de logística de envíos para poder responder ¿cuándo llega?

# Descripción del dataset

Datos: El conjunto de datos seleccionado para realizar el práctico corresponde a un muestreo aleatorio no uniforme de 500.000 envíos de MercadoLibre. Estos envíos fueron realizados en Brasil en el período comprendido entre Octubre de 2018 y Abril de 2019 (las fechas han sido modificadas y adaptada a un período de tiempo diferente, conservando el día de la semana y considerando los feriados correspondientes). Los datos comprenden variables tanto categóricas como numéricas.

El dataset presenta la siguiente información:

* Sender\_state: Estado de Brasil de donde sale el envío.
* Sender\_zipcode: Código postal (de 5 dígitos) de donde sale el envío.
* Receiver\_state: Estado de Brasil a donde llega el envío.
* Receiver\_zipcode: Código postal (de 5 dígitos) a donde llega el envío.
* Shipment\_type: Método de envío (normal, express, super).
* Quantity: Cantidad de productos en un envío.
* Service: Servicio del correo con el cual se realizó un envío.
* Status: Estado del envío (set: listo para ser enviado, sent: enviado, done: entregado, failed: no entregado, cancelled: cancelado).
* Date\_created: Fecha de creación del envío.
* Date\_sent: Fecha y hora en que se realizó el envío (salió del correo).
* Date\_visit: Fecha y hora en que se entregó el envío al destinatario.
* Shipment\_days: Días hábiles entre que el envío fue enviado (salió del correo) y que fue entregado.

# Objetivo

El objetivo de este práctico es aprender a leer y analizar el dataset para extraer información útil sobre el problema a resolver. Se espera que se gane conocimiento en el manejo de librerías y metodologías de visualización para generar gráficas de los datos. Otro objetivo es calcular estadísticos que faciliten, junto a los gráficos, dar respuestas a preguntas propias de la problemática. Por último, se busca desarrollar habilidades de comunicación de la información obtenida a partir de los datos de manera clara y sencilla.

# Método

A partir de lo estudiado en las clases teóricas y prácticas de la materia “Análisis y visualización de datos”, realizar un informe en formato de notebook o interactivo, en el cual se describen los aspectos de mayor relevancia del dataset. En la misma se deben incluir las gráficas y los estadísticos que sean pertinentes para responder a las preguntas indicadas a continuación junto con el análisis y las respuestas propiamente dichas (además de cualquiera otra información extra que se considere de relevancia sobre la problemática):

* 1 ¿Qué tipo de variables se encuentran presentes en el dataset? Realizar una breve descripción estadística de las mismas, determinar si existen outliers y valores incorrectos o no esperados.
* 2 ¿Cómo es la distribución de zipcodes de origen y destino de un envío? ¿Está concentrada en algún valor? ¿Y para el resto de las variables?
* 3 ¿Es la distribución del número de días que demora un envío normal? ¿Adquiere significado la media y desviación estándar? ¿Qué tipo de estadísticos podrían ser relevantes en este caso? ¿Hay alguna otra distribución a la cual se asemeje más?
* 4 ¿Cuál es la probabilidad de que el destinatario de un envío sea del estado de Minas Gerais (MG)? ¿Cuál es la probabilidad de que dicho envío (que llega a Minas Gerais) sea realizado mediante el servicio 1?
* 5 ¿Cuál es la probabilidad de que un envío sea realizado con el método super express? ¿Y la de que un envío realizado con dicho método llegue al estado de Minas Gerais?
* 6 ¿Existe correlación entre el número de días que tarda un envío y el destinatario del mismo? ¿Y entre el número de días y la cantidad de productos en un paquete? ¿Con qué otra variable está correlacionado el número de días que tarda un envío? Analizar la independencia de las variables.
* 7 Además del cálculo de los estadísticos correspondientes, sería interesante observar un gráfico que muestre la información estadística (por ejemplo, la media) y la presencia de outliers del número de días que tardan los envíos que se realizan a los diferentes estados de Brasil. ¿Cuál es el estado que, en promedio, tiene una mayor demora? ¿Cuál es la distancia geográfica de San Pablo a dicho estado? ¿Y al de menor?
* 8 ¿Cuál es la distribución de la entrega de envíos *por hora del día* para cada uno de los estados de Brasil? Realizar un único gráfico donde se muestre esta información.
* 9 Bonus (no obligatorio): Proponer y discutir la aplicación de al menos uno de los contenidos desarrollados en la clase de “Inferencia Bayesiana” al conjunto de datos y problemática aquí presentada.

Esta comunicación debe estar dirigida para un público técnico pero que desconoce los aspectos propios del problema a resolver (por ejemplo, sus compañeros de clase). Se evaluará, principalmente, la claridad del mensaje presentado, el uso de las herramientas y los conceptos de visualización desarrollados en las clases, y el cálculo de los estadísticos apropiados para el problema.

# Estructura del informe

El informe debe contar con la estructura propia de un reporte de un experimento científico. Esto implica que debe tener un objetivo claro, una introducción a la problemática a resolver en dicho informe (no únicamente al problema general), una descripción de los datos a emplear, el desarrollo propiamente dicho del experimento y las conclusiones que se obtuvieron.

En el informe se deberá brindar una descripción del dataset suministrado (columnas, tipo de variables, valores extremos, etc.), las visualizaciones realizadas que sean pertinentes para la resolución del práctico, un análisis de los gráficos y estadísticos calculados, el análisis y las respuestas a las preguntas indicadas anteriormente, y las conclusiones.

# Entrega

La entrega del informe final será antes del día 17 de Mayo, con una muestra previa de avance el día 10 de Mayo. El notebook donde se realicen los cálculos y gráficos debe encontrarse en un repositorio al cual se pueda acceder.