Informe Taller 3 – Análisis Bivariado

Elaborado por:
Holman Sanchez
Esteban Oliveros
Jesús Córdoba

Contexto y preparación de datos

- **Fuente:** base de obras públicas (contratos) cruzada con facturas.
- **Limpieza previa esencial:** estandarización de **fechas** (parseo y normalización de formatos para poder cruzar y calcular duraciones), eliminación/tratamiento de **NA** en variables clave (valor_contrato, dias_fin_ejec, bandera, municipio_entidad, porcentaje) y depuración de **outliers manifiestos** por criterio de negocio (p. ej., registro con porcentaje ≈ 96.000% eliminado por inconsistencia).
- **Variable derivada relevante:** porcentaje = pagos acumulados / valor estimado del contrato × 100. Se usa para diagnosticar (sub/ sobre) ejecución financiera.

1) Selección de variables y justificación

Par principal de análisis (numérica vs numérica)

- **valor_contrato** (monto del contrato)
- **dias_fin_ejec** (duración hasta el fin de ejecución)

Justificación: Estos dos indicadores miden la magnitud financiera y la magnitud temporal de cada obra. Entender su relación permite evaluar si los proyectos más costosos tienden a requerir más tiempo insumo clave para planificación, programación y control de riesgos.

Variables de apoyo

- **bandera** (0/1: presencia de sobrecostos > 30%): permite evaluar si el riesgo de sobrecostos cambia con el tamaño y la duración.
- **municipio_entidad** (categórica): permite observar **heterogeneidad territorial** en la ejecución (aplicado con porcentaje).

2) Análisis bivariado (resultados con base en el notebook)

2.1) Dispersión: valor_contrato vs dias_fin_ejec (escala log–log)

- **Visualización:** scatterplot en log–log para reducir sesgo por colas largas; puntos coloreados por bandera.
- **Patrón observado: tendencia positiva** clara: a medida que crece el valor del contrato, aumentan los días de ejecución. La nube es amplia (no hay proporcionalidad 1:1), pero la pendiente positiva es consistente a lo largo de varios órdenes de magnitud.

- **Correlación (Pearson): r = 0.31** → **relación positiva moderada**.
    - **Interpretación:** ~**9.6%** de la variabilidad de dias_fin_ejec puede explicarse linealmente por valor_contrato ($R^2 \approx 0.31^2$). Es una relación real pero **no determinística**: influyen otros factores (alcance, sector, modalidad, gestión, entorno local).

2.2 Riesgo de sobrecostos: valor_contrato vs bandera

- **Visualización:** scatter valor_contrato (eje x, log) vs bandera (0/1).
- **Patrón observado: no hay estratificación** de sobrecostos por tamaño: aparecen banderas tanto en contratos pequeños como grandes.
- **Correlaciones (point-biserial / Pearson con binaria):**
    - bandera vs valor_contrato: **r ≈ -0.03**
    - bandera vs dias_fin_ejec: **r ≈ -0.06**
- **Interpretación:** efectos **prácticamente nulos**; el valor o la duración por sí solos **no explican** la aparición de sobrecostos. Se requiere integrar más variables (modalidad, sector, tipo de interventoría, proveedor, régimen, fuente de recursos, etc.).

2.3 Heterogeneidad territorial: porcentaje vs municipio_entidad

- **Visualización:** boxplots por municipio.
- **Patrón observado: dispersión heterogénea** entre municipios:
    - Casos como **Tuluá** y **Palmira** muestran **IQR más amplio y colas superiores** largas (más variabilidad y presencia de sobre-ejecuciones).
    - Municipios como **Andalucía** o **Calima** presentan **concentración** alrededor de 100% (ejecución más estable).
- **Lectura sustantiva:** hay **diferencias territoriales** en comportamiento de ejecución financiera (y por ende en riesgo operativo/administrativo). Esto sugiere priorización territorial para auditorías y acompañamiento técnico.

2.4 Matriz de correlación (resumen)

- Heatmap entre dias_fin_ejec, valor_contrato y bandera:
    - valor_contrato – dias_fin_ejec: **+0.31** (moderada, ya discutida).
    - bandera con ambas: **cercanas a 0** (nulas/irrelevantes).

3) Interpretación integrada

1. **Costo vs tiempo:** existe una relación positiva moderada. Proyectos más costosos tienden a durar más, pero la gran dispersión indica que la gestión y el contexto importan: hay contratos de similar valor con duraciones muy distintas (diferencias en alcance, complejidad técnica, condiciones del sitio, logística, permisos, etc.).

2. **Sobrecostos transversales:** la presencia de banderas no se concentra en un rango de valores ni duraciones; es transversal. Esto sugiere que los sobrecostos no son una "propiedad" del tamaño, sino de cómo se contrata, supervisa y ejecuta (modalidad, interventoría, gobernanza, capacidades del contratista, cambios de alcance, adiciones, etc.).
3. **Brecha territorial:** las diferencias entre municipios son relevantes. Donde hay mayor variabilidad en porcentaje, hay más incertidumbre en la ejecución. Ello puede orientar focalización de control (auditorías, alertas tempranas, asistencia técnica).

4) Conclusiones

1. **Existe una relación positiva y moderada entre el monto del contrato y su duración** ($r \approx 0.31$). Los proyectos más grandes tienden a extenderse más, pero el valor no es suficiente para predecir el tiempo: se requieren variables de gestión y contexto para mejorar la explicación.
2. **Los sobrecostos (bandera) no están asociados de forma significativa ni al** valor ni al tiempo (correlaciones cercanas a cero). Por lo tanto, no se recomienda usar únicamente estas variables para anticipar banderas de sobrecostos.
3. **Hay heterogeneidad territorial** en la ejecución financiera (porcentaje), con municipios que exhiben mayor dispersión y colas superiores. Esto justifica estrategias de vigilancia diferenciadas y la priorización de auditorías por territorio.
4. **Implicación práctica:** un sistema de monitoreo predictivo debe incorporar variables de proceso (modalidad de contratación, sector, régimen, tipo de interventoría, historial del contratista, enmiendas/adiciones, hitos de obra, etc.) y variables territoriales para captar los verdaderos determinantes del riesgo de sobrecosto y retraso.
5. **Próximos pasos recomendados:**

- Construir **modeloslogísticos** para bandera incorporando: **modalidad**, **sector**, **nivel_entidad**, **municipio**, **valor_categoria**, y variables de calendario (año, trimestre).
- Evaluar ANOVA (según supuestos) para porcentaje entre municipios y sectores.